XV WORLD FORESTRY CONGRESS
Building a Green, Healthy and Resilient Future with Forests
2–6 May 2022 | Coex, Seoul, Republic of Korea

Coupling machine learning and forest simulations to promote the applicability of long-term forest projections under climate change perspective

Irina Cristal[1,2,3], Blas Mola[4], Jose Ramon Gonzales Olabarria[1] , Jordi Garcia Gonzalo[1]

[1][Forest Science and Research Center of Catalonia, Spain]
[2][University of Lleida, Spain]
[3][Forest Bioengineering Solutions, Spain]
[4][University of Western Finland, Finland]

Abstract

Projecting forest dynamics is the foundation for sound decision support in adaptive forest management. However, due to their complexity, many forest modeling techniques addressing global changes in terrestrial ecosystems are limited to scientific applications. Integrating conventional research and artificial intelligence technologies has the potential to bridge research and practical use.

In this study, we propose a Machine Learning (ML) framework that facilitates the implementation of long-term forest projections under climate change scenarios. Our approach combines ML and forest simulations based on process-based models to project forest dynamics. The goal is to leverage the complementary strength of process-based and state-of-the-art ML models to improve predictions at a reduced computational cost. We use environmental data and periodic field measurements at a national scale to train ML models to predict forest growth. By integrating process-based simulations we investigate how the additional variables can improve the prediction accuracy.

The proposed hybrid ML framework identifies forest dynamics processes and drivers across spatial and temporal scales, contributing at many levels to the climate change adaptation: from increasing awareness of the climate-induced hazards to enhancing education and assisting in sustainable natural resource management and planning.

Keywords: adaptive forest management, climate change, forest growth modelling, machine learning

Introduction

Knowledge about forest systems and their future development is essential when it comes to sustainable management of forest resources under changing climate. Forest growth models are key components in understanding and projecting forest changes over temporal and spatial scales (Weiskittel et al. 2011, Terrades 2005, Larocque 2015, Tenzin et al. 2017). As compared to field studies, modelling techniques give the opportunity to analyze the possible impact of different alternative practices for a wide range of environmental settings. During the past decades, climate change has imposed a shift in management objectives from focusing on a continuous revenue, to increasing forest resilience to global changes. As a result, a plethora of models explaining forest responses to new threats have emerged. Two main groups of models exist to describe forest dynamics: theory-based and data-driven (Korzukhin et al., 1996, Mäkelä et al. 2000). Approaches based on eco-physiological and mechanistical processes in plants are called process-based (PB) or mechanistical, and use mathematical formulations to describe the underlying processes and their

interactions (Adams et al. 2013). Computer simulations of the PB models are able to predict vegetation dynamics shifts under climate change scenarios and can assist in better ecosystem management (García-Gonzalo et al. 2007). However, these models are complex and difficult to parameterize. The parameterization, usually field-based, can introduce uncertainty at multiple levels related to the number of processes involved (Adams et al. 2013, Olsson and Jönsson 2014). These models tend to be species specific and spatially restricted. As a result, process-based models are rarely used in practical implementations of forest management and planning (Schuwirth et al 2019).

On the other hand, data-driven models use past observations to infer the relationships between forest attributes (Korzukhin et al. 1996, Adams et al. 2013). Advances in machine learning (ML) today manifested into data-driven approaches with increased predictive capability (e.g. Ashraf et al 2015, Jevšenak and Skudnik 2021). ML takes advantage of a large amount of data to model accurate predictive models without considering domain knowledge. However, field measurements at large scales are costly, and the currently available data are not sufficient for developing unbiased and reliable models solely on past observations (Adams et al. 2013, Pukkala et al. 2021).

We argue that coupling machine learning with PB modelling can promote multiple aspects of the usability of forest models in forest management decision making. The added value of ML is related to the flexibility in applying the models across spatial and temporal scales, the lowered uncertainty of prediction, the quantified probability, and the reduced computational time. By integrating process-based simulations, the ML model can account for missing data which are needed to increase the predictive accuracy for different environmental conditions in the future.

With the present work we aim at improving the applicability of forest simulations in decision making. We developed a hybrid approach by combining the reliability of a process-based model and the predictive accuracy of machine learning. We are testing our hypothesis on a dataset extracted from the periodical field measurements of the Spanish National Forest Inventory.

Methodology

Modelling approach

In this study we used measurements of individual trees from Spanish National Forest Inventory (NFI), climate data, and simulation of forest dynamics to train a Random Forest (RF) model to predict forest growth at an individual tree level (Fig. 2). Both, historical and projected tree characteristics, coupled with additionally computed variables were employed in training and validation of the ML-based diameter increment model.

Study area

By choosing Forema-DST to project forest dynamics, we limited our study area to Catalonia, north-east Spain (Fig. 1). Elevation of the area ranges from 0 to 3000 meters above sea level, creating climate belts which highly contribute to the climate and weather variability and the heterogeneity of forest cover. Forests in Catalonia account for 60% of the territory, of which pines and oaks are the dominant tree species. One of the most important species in Catalonia (1st regarding growing stock and harvested volume, and 2nd in covered surfaces) is *Pinus sylvestris*, a species adapted to a wide

range of conditions, and abundant in the Catalan mountain ranges within altitudes between 800 and 1600 m.a.s.l (Gracia et al., 2004).

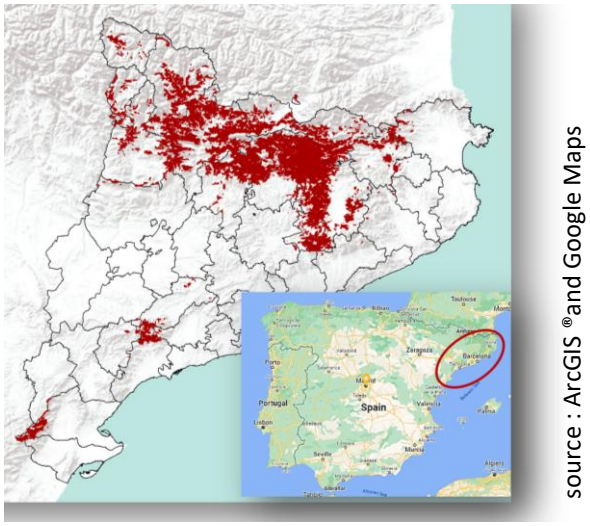

Fig. 1: Study area: administrative borders of Catalonia and the reference map of Spain. Red points indicate the location of *P. sylvestris* plots retrieved from the Spanish NFI and used in this study.

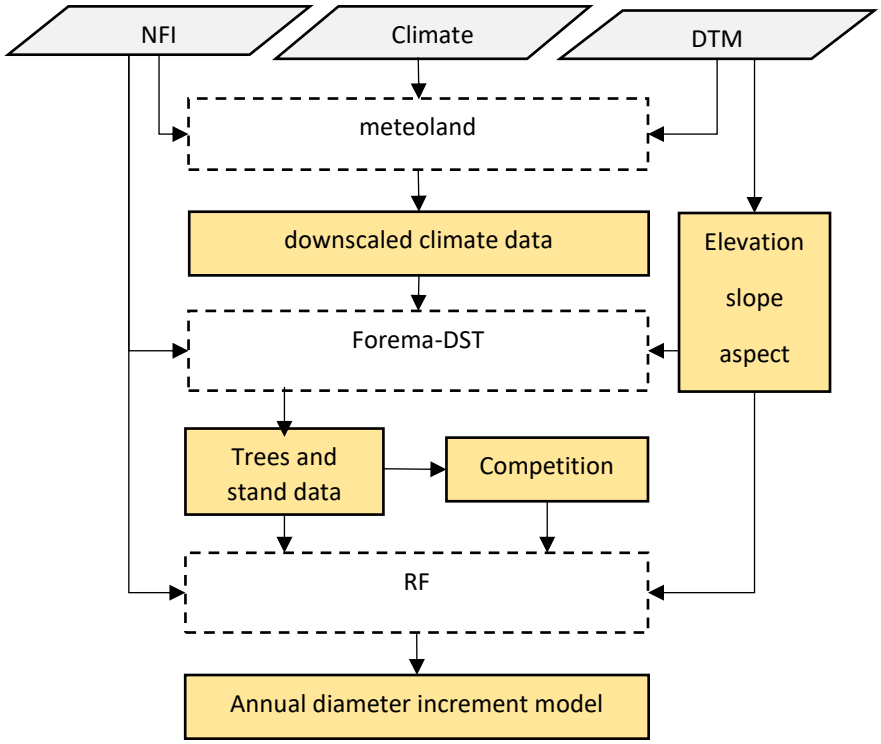

Fig. 2: Flow diagram of the hybrid ML-based approach. Climate data are downscaled on the plot coordinates and further processed together with tree and stand data from the NFI to obtain future projections of the plots. Finally, simulated timeseries coupled with additional data from NFI and extracted features from DTM are used to generate the training dataset for the RF model. As a result, the model predicts annual diameter increment.

Data preparation

Candidate predictors for the ML-based model were based on literature and previous studies (e.g. Pukkala et al. 2021, Ashraf et al. 2015). All explanatory variables can be grouped in (i) tree size variables, (ii) site characteristics, (iii) competition indices and (iv) climate data (Table 1). NFI stand and tree level data were used to prepare the input for the process-based simulations, as well as for the ML-based model. Additional data were either extracted from different sources (e.g. digital terrain model - DTM) or were computed from the simulation outputs (e.g. basal area)

Table 1: Main explanation variables used in this study.

variable group	variable	Initial source	use
Tree size	DBH	NFI	RF
	Basal Area (BA)	NFI	RF
	Height	NFI	RF
Site characteristics	Elevation	DTM	RF/ Forema-DST
	Slope	IFN	RF/ Forema-DST
	Aspect	DTM	RF/ Forema-DST
	Organic matter	IFN	RF
Competition	BA in larger trees	Forema-DST	RF
	Dominant height	Forema-DST	RF
Climate	Precipitation	meteoland	Forema-DST
	Temperature	meteoland	Forema-DST

Spanish National Forest Inventory

Permanent plots in Catalonia were established between 1986 and 1996 during the 2nd NFI. The 3rd NFI took place between 2000 and 2001. Fifty-two forest types were reported, covering an area of 32.114 km^2, along an elevation gradient ranging from 0 to 2300 m above sea level (Gonzalez et al. 2007). In the Spanish NFI trees are measured in a variable radius plot, as a function of their Diameter at Breast Height (DBH). We extracted pure *Pinus sylvestris* plots measured in the 2nd and the 3rd NFI from the NFI databases (ICONA 1995, DGCN 2005). Both, tree and stand level information was used in the platform Forema-DST (Cristal et al. 2019) to simulate forest dynamics.

Forest dynamics simulation

Forema-DST combines a process-based gap model (i.e. Sortie-ND, Canham et al. 2001) parameterized for four species in Catalonia (Ameztegui et al. 2015) with empirical models of ecosystem services (Cristal et al. 2019). The simulator takes as input stand and trees characteristics, and climate data. Timeseries of precipitation and temperature were extracted from the weather stations in Catalonia and extrapolated to the study area stands using the R package “meteoland” (De Caceres et al. 2015).

The output of the forest growth simulator embodies tree level data per each year of the simulation, namely, species, DBH, height, light, crown radius, crown depth, and tree coordinates.

Using these outputs, additional variables were calculated.

- Diameter increment (DI), as the diameter difference in two consecutive years.

- Stand Basal Area (BA), as the sum of BAs of all the individual trees in the stand.
- Dominant Height (DH), as the mean height of the largest 100 trees in the stand.
- Basal Area in Larger trees (BAL), as the sum of BAs of all the trees larger than the referenced one (Wykoff 1990).

A total of 23 predictors and nearly 3 million observations of trees at different timesteps were recruited to train the random forest regression model to predict diameter increment.

Random Forest

Random forest is a popular machine learning algorithm, mainly because of the clarity in its formulation and the ability to give reliable results (Bergman 2001, Ho 1995). It falls under the supervised learning problem that operates by constructing multiple decision trees (Ho 1995). The algorithm randomly selects observations and a subset of explanatory variables from the training dataset to build decision trees. A decision tree itself is a learning algorithm, which derives relations between predictors and target variables by splitting the dataset according to the order of selected predictors. The order of predictors, or explanatory variables, is selected with the goal to reduce variability in observations (entropy). Final decision (prediction) is made by aggregating ensemble of these decision trees, known as "bagging" (Bergman 2001), either on the majority of votes in the case of classification, or the mean values in the case of regression trees (Ho 1998). The process of randomly selecting observations and explanatory variables, prevents the trees from overfitting. The main idea of ensemble models is that combining multiple learning algorithms increases the accuracy of the predictions.

Based on previous studies, RF algorithm outperformed other ML approaches in predicting growth rates based on timeseries data (Shahhosseini et al. 2020, Jevšenaka and Skudnika 2021).

We applied Forest based classification and regression trees implementation in ArcGIS Pro to model the diameter increment in individual trees. The training dataset consisted of nearly 3 million of observations, of which 10% was left for validation. In the model parameters we configured the number of random decision trees per forest, the number of randomly selected variables tree, and the minimum size of nodes.

Results

ML-based diameter increment model had 11 important predictors showed in Table 4. Indicator variable BAL affected diameter increment the most (up to 37%), implying that competition is an important factor in *P. Sylvestris* growth. The model captured the variation of the explanatory variables at 84% in the training dataset, and 83,8% in the validation dataset (Table 3). The validation was performed for 100 iterations. The distribution of R2 value (Figure 3) shows that most of the models have high predictive accuracy. The accuracy increased with the increased number of trees of the model (Table 2). Random Forest combines different predictors (decision trees) to make predictions. Thus, increasing the number of trees result in a more robust model.

Table 1. Model characteristics

Number of Trees	150
Leaf Size	5
Tree Depth Range	5-5
Mean Tree Depth	5
% of Training Available per Tree	100
Number of Randomly Sampled Variables	5
% of Training Data Excluded for Validation	10

Table 2. Model Out of Bag Errors

Number of Trees	75	150
MSE	0.004	0.003
% of variation explained	82.824	83.134

Table 3: Regression Diagnostics

Metrics	Training*	Validation**
R-Squared	0.840	0.838
p-value	0.000	0.000
Standard Error	0.001	0.002

*Predictions for the data used to train the model compared to the observed categories for those features.
**Predictions for the test data (excluded from model training) compared to the observed values for those test features.

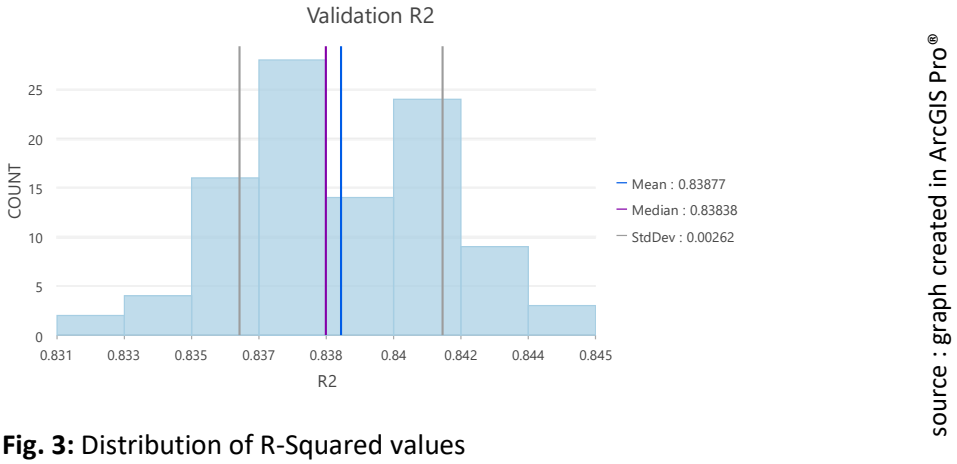

Fig. 3: Distribution of R-Squared values

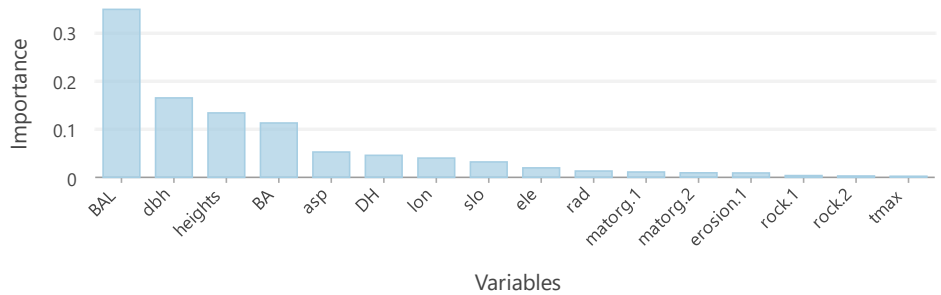

Fig. 4: Distribution of variables of importance. Basal area in largest trees is the stronger predictor in the RF model, followed by tree size variables.

Table 4: Variables of importance

Variable	Importance	%
BAL	1591.72	37
height	584.32	14
BA	567.29	13
DBH	526.72	12
latitude	245.86	6
dominant height	167.48	4
aspect	164.18	4
longitude	142.58	3
slope	96.40	2
elevation	71.37	2
radiation	43.72	1
organic matter	40.68	1

Discussion

Both ML and PB modeling have been separately studied in the context of forest growth modelling. PB models are widely used in projecting the impact of climate change on forests. However, despite their wide application in the scientific context, these models are rarely used in practice. This fact may be related to the induced complexity in the models and the large computational requirements of their simulations. In contrary, ML doesn't require domain knowledge, and can make accurate predictions based on big number of past observations. Nevertheless, the currently available forest monitoring data do not allow to develop reliable models, especially for the changing climate conditions. The proposed hybrid framework combines the ability to capture future changes based on a sound physiological theory of process-based modeling and the prediction accuracy of the data-driven machine learning approach. We tested our hypothesis by training a random forest regression model on tree level data obtained from Spanish NFI and their projection using the platform Forema-DST.

The R2 of the model on both training and validation datasets was ranging between 0.82 and 0.84 depending on the number of decision trees employed and the depth of the trees. The RF model outperformed previously reported studies that employed only observation data. The most important explanatory variable was the competition index BAL, followed by tree size variables.

Conclusions

Concluding, the proposed framework revealed a great potential of using ML to overcome the limitations of PB models in practice: once trained, the model can run the simulations at a minimal computational time, and without domain knowledge of the PB modelling (e.g. parameterization). The accuracy of the predictions increased compared to the ML models based solely on past observations. However, further work is needed to improve the model applicability by employing more species and management alternatives.

Acknowledgements

This work was supported by Microsoft AI for Earth program. Many thanks to Lena Vilà Vilardell for providing valuable support with the R package “meteoland”.

References

- **Periodicals:**

  Adams H, Williams A, Xu C, Rauscher S, Jiang X, McDowell N. 2013. Empirical and process-based approaches to climate-induced forest mortality models. Frontiers in Plant Science, 4: 438

  Ameztegui A, Coll L, Messier C. 2015. Modelling the effect of climate-induced changes in recruitment and juvenile growth on mixed-forest dynamics: The case of montane–subalpine Pyrenean ecotones. Ecological Modelling, 313: 84-93

  Ashraf MI, Meng FR, Bourque CPA, MacLean DA. 2015. A novel modelling approach for predicting forest growth and yield under climate change. PloS one 10 (7): e0132066

  Breiman L. 2001. Random Forests. Machine Learning. 45 (1): 5–32.

  Canham CD, Coates KD, Bartemucci P, Quaglia S. 1999. Measurement and modeling of spatially explicit variation in light transmission through interior cedar-hemlock forests of British Columbia. Can. J. For. Res., 29: 1775–1783.

  Cristal I, Ameztegui A, González-Olabarria JR, Garcia-Gonzalo J. 2019. A Decision Support Tool for Assessing the Impact of Climate Change on Multiple Ecosystem Services. Forests, 10(5): 440

  De Caceres M, Martin-StPaul N, Turco M, Cabon A, Granda V. Estimating daily meteorological data and downscaling climate models over landscapes. 2018. Environ. Model. Softw. 108: 186–196.

  González JR, Trasobares A, Palahí M, Pukkala T. 2007. A fire probability model for forest stands in Catalonia (north-east Spain). Annals of forest science. 64 (7): 733-742.

  Ho TK. 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16.

  Ho TK. 1998. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.

  Jevšenaka J, Skudnika M. 2021. Random forest model for basal area increment predictions from national forest inventory data. Forest Ecology and Management 479: 118-601

Korzukhin MD, Ter-Mikaelian MT, Wagner RG. 1996. Process versus empirical models: which approach for forest ecosystem management? Canadian Journal of Forest Research, 26: 879 - 887.

Mäkelä A, Landsberg J, Ek AR, Burk TE, Ter-Mikaelian M, Agren GI, Oliver CD, Puttonen P. 2000. Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation. Tree Physiology, 20(5_6): 289-298.

Olsson C, Jönsson AM. 2014. Process-based models not always better than empirical models for simulating budburst of Norway spruce and birch in Europe. Global Change Biology, 20 (11): 3492-3507.

Pukkala T, Vauhkonen J, Korhonen KT and Packalen T. Self-learning growth simulator for modelling forest stand dynamics in changing conditions. 2021. Forestry: An International Journal of Forest Research, 94 (3): 333–346.

Schuwirth N, Borgwardt F, Domisch S, Friedrichs M, Kattwinkel M, Kneis D, Kuemmerlen M, Langhans S, Martínez-López J, Vermeiren P. 2019. How to make ecological models useful for environmental management. Ecological Modelling, 411: 108784.

Shahhosseini M, Hu G, Huber I, Archontoulis SV. 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports, 11: 1606

Tenzin J, Tenzin K., Hasenauer H. 2017. Individual tree basal area increment models for broadleaved forests in Bhutan. Forestry, 90: 367–380.

Terradas, J. 2005. Forest dynamics: A broad view of the evolution of the topic, including some recent regional contributions. Sistemas y recursos forestales 14(3): 525-537.

Wykoff W. R. 1990. A basal area increment model for individual conifers in the northern Rocky Mountains. Forest Science, 26: 1077-1104.

- **Books:**

DGCN, Tercer Inventario Forestal Nacional (1997–2007) Cataluña: Barcelona, Ministerio de Medio Ambiente, Madrid, 2005

Garcia-Gonzalo J, Zubizarreta-Gerendiain A, Kellomäki S, Peltola S. 2017. Managing Forest ecosystems: the challenge of climate change. Springer. 277-298 pp.

Gracia C, Burriel JA, Ibàñez JJ, Mata T, Vayreda J. 2004. Inventari Ecològic i Forestal de Catalunya. CREAF, Bellaterra. 184 pp.

ICONA, Segundo Inventario Forestal Nacional (1986–1995), Cataluña: Barcelona, Madrid, 1993

Killham K. 1994. Soil ecology. Cambridge: Cambridge University Press, 242 pp.

Larocque, G.R. 2016. Ecological Forest Management Handbook. CRC Press. 604 pp.

Weiskittel, Aaron R.; Hann, David W.; Kershaw, John A. Jr; and Vanclay, Jerome K. 2011. Forest growth and yield modeling. Hoboken, NJ: Faculty and Staff Monograph Publications. 415 pp.