



## CROMES - A fast and efficient machine learning emulator pipeline for gridded crop models

**Christian Folberth**<sup>1</sup>, Artem Baklanov<sup>2</sup>, Nikolay Khabarov<sup>2</sup>, Thomas Oberleitner<sup>1</sup>, Juraj Balkovic<sup>1</sup>, and Rastislav Skalsky<sup>1</sup>

<sup>1</sup>Biodiversity and Natural Resources Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

<sup>2</sup>Advancing Systems Analysis Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Global gridded crop models (GGCMs) have become state-of-the-art tools in large-scale climate impact and adaptation assessments. Yet, these combinations of large-scale spatial data frameworks and plant growth models have limitations in the volume of scenarios they can address due to computational demand and complex software structures. Emulators mimicking such models have therefore become an attractive option to produce reasonable predictions of GGCMs' crop productivity estimates at much lower computational costs. However, such emulators' flexibility is thus far typically limited in terms of crop management flexibility and spatial resolutions among others. Here we present a new emulator pipeline CROp model Machine learning Emulator Suite (CROMES) that serves for processing climate features from netCDF input files, combining these with site-specific features (soil, topography), and crop management specifications (planting dates, cultivars, irrigation) to train machine learning emulators and subsequently produce predictions. Presently built around the GGCM EPIC-IIASA and employing a boosting algorithm, CROMES is capable of producing predictions for EPIC-IIASA's crop yield estimates with high accuracy and very high computational efficiency. Predictions require for a first used climate dataset about 45 min and 10 min for any subsequent scenario based on the same climate forcing in a single thread compared to approx. 14h for a GGCM simulation on the same system.

Prediction accuracy is highest if modeling the case when crops receive sufficient nutrients and are consequently most sensitive to climate. When training an emulator on crop model simulations for rainfed maize and a single global climate model (GCM), the yield prediction accuracy for out-of-bag GCMs is  $R^2=0.93-0.97$ ,  $RMSE=0.5-0.7$ , and  $rRMSE=8-10\%$  in space and time. Globally, the best agreement between predictions and crop model simulations occurs in (sub-)tropical regions, the poorest is in cold, arid climates where both growing season length and water availability limit crop growth. The performance slightly deteriorates if fertilizer supply is considered, more so at low levels of nutrient inputs than at the higher end.

Importantly, emulators produced by CROMES are virtually scale-free as all training samples, i.e., pixels, are pooled and hence treated as individual locations solely based on features provided without geo-referencing. This allows for applications on increasingly available high-resolution

climate datasets or in regional studies for which more granular data may be available than at global scales. Using climate features based on crop growing seasons and cardinal growth stages enables also adaptation studies including growing season and cultivar shifts. We expect CROMES to facilitate explorations of comprehensive climate projection ensembles, studies of dynamic climate adaptation scenarios, and cross-scale impact and adaptation assessments.