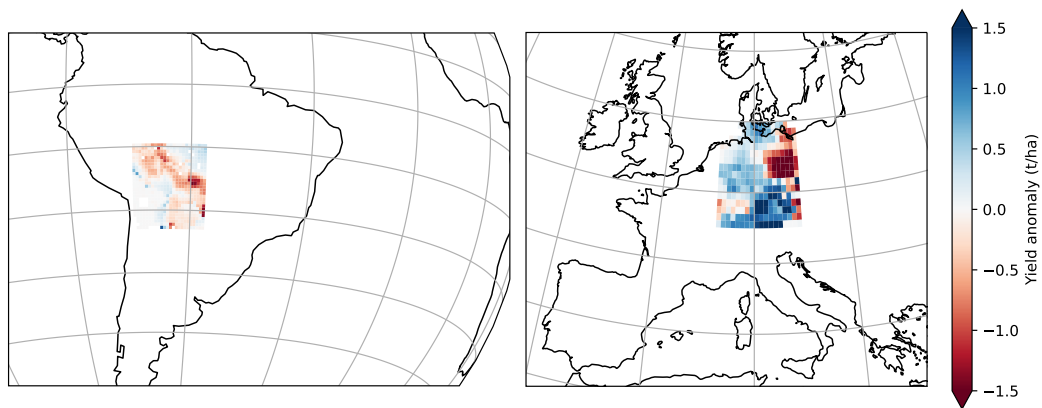# Supplemental Material

Figure S1: Maps showing the spatial regions withheld from the training set, thereby making up the 'unseen regions' for model evaluation. Color shows the yield anomalies (the target variable) for the 15th harvest in the dataset. The first region lies between -22.8 and -9.6 ° latitude and -69.6 and -57.4 ° longitude. The second region lies between 47.3 to 55.1 ° latitude and 5.9 to 15.0 ° longitude.
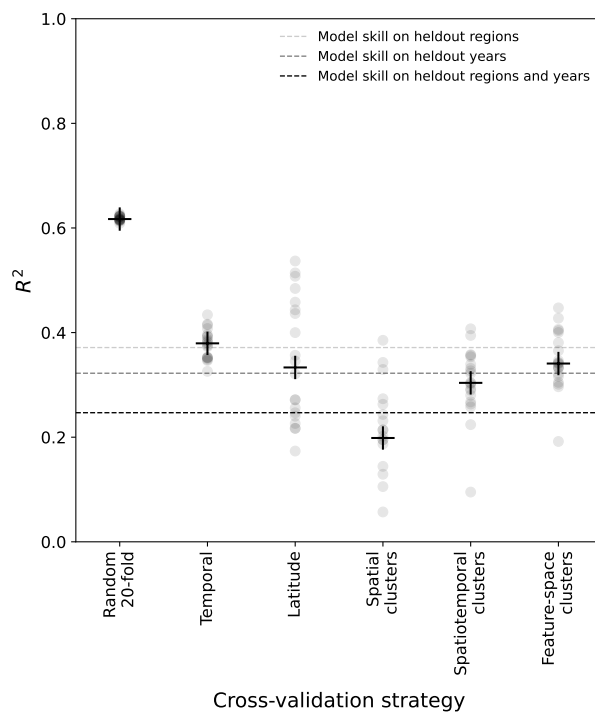

Figure S2: Model skill as evaluated using various CV strategies on the training set, for pDSSAT model output. Each point represents the performance as measured using one CV test fold and crosses represent the median scores over all twenty folds. Horizontal dashed lines denote, for comparison, the model skill (trained on the entire training set) on the heldout test sets.
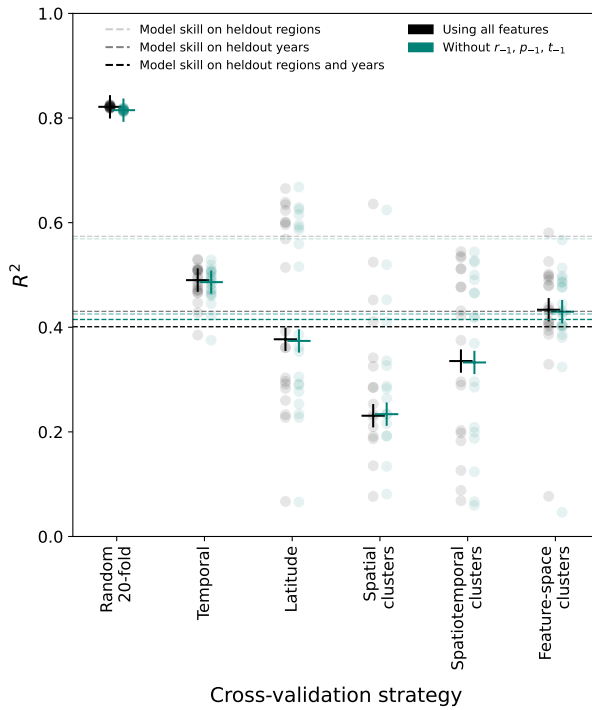
Figure S3: Model skill as evaluated using various CV strategies on the training set and on the three unseen test sets, using all features and without features r_{-1}, p_{-1} and t_{-1}.
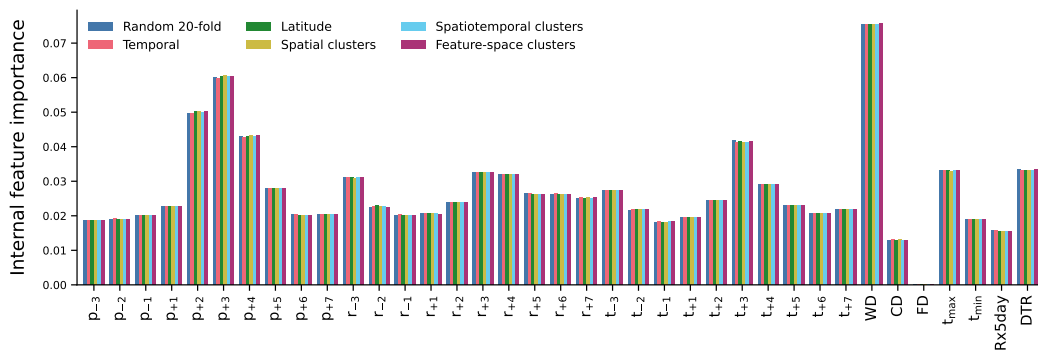

Figure S4: Median impurity-based feature importances of the random forest models across 20 CV folds, using each CV strategy. Impurity-based feature importances are equal to the mean decrease of impurity using that feature within each decision tree of the forest, as computed during training.
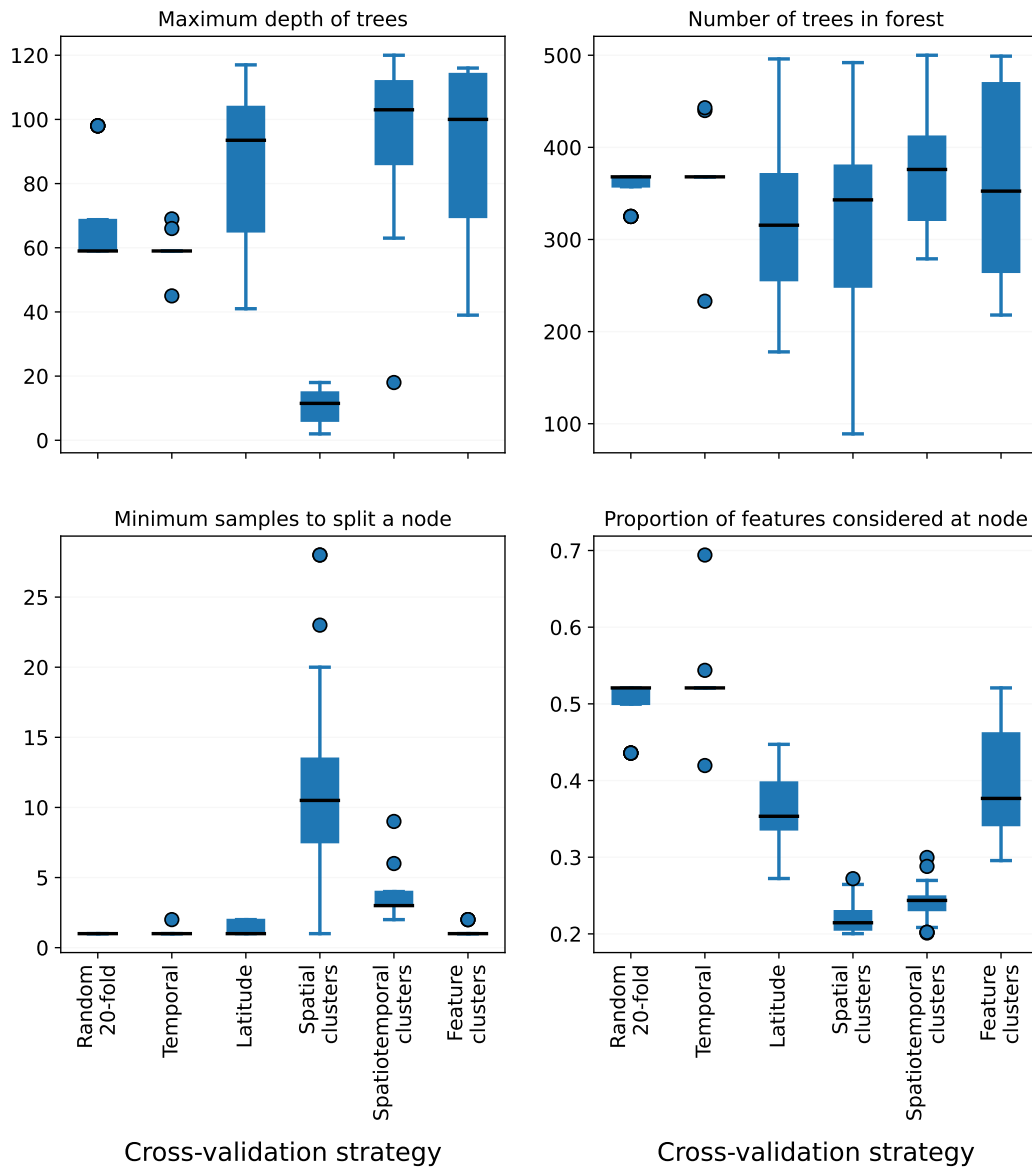
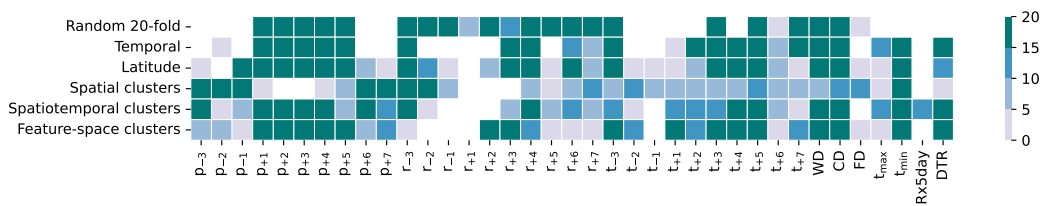Figure S5: Hyperparameters selected using the various CV methods.

Figure S6: Features selected using the various CV methods. Colour denotes the number of CV folds for which a feature was selected. White fields denote features which were not selected for any training folds using that CV strategy.
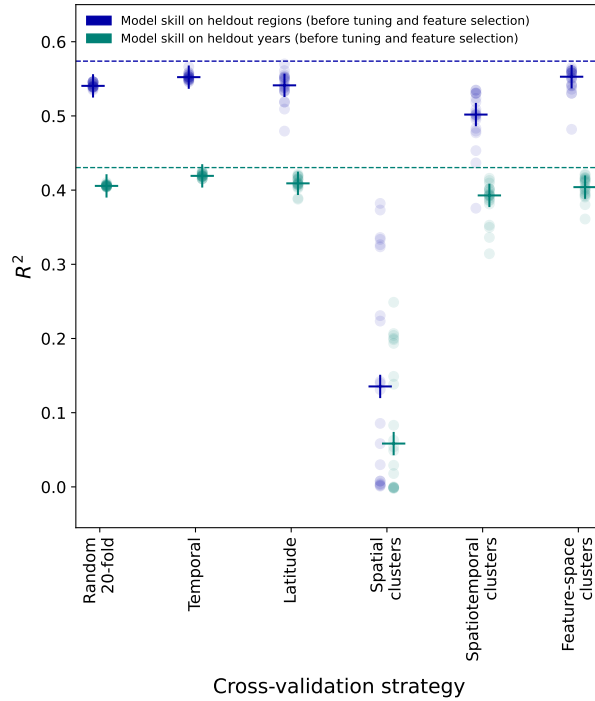
Figure S7: Model skill after hyperparameter-tuning and feature-selection was conducted using six different CV strategies, as measured on unseen years or unseen regions. Each point represents the performance of one resulting model which was trained on one CV training fold, and crosses denote the median performance across the twenty resulting models.

Figure S8: Ranked feature importances of random forest models, for which nested CV is used for hyperparameter-tuning and feature-selection steps, before importances are evaluated on outer CV test folds, for the 37 climatic features. Shades denote the median score across the twenty test folds and size of circles gives a qualitative measure of the range in feature importances across the CV test folds. White fields are features which are not selected for any folds using that CV strategy.
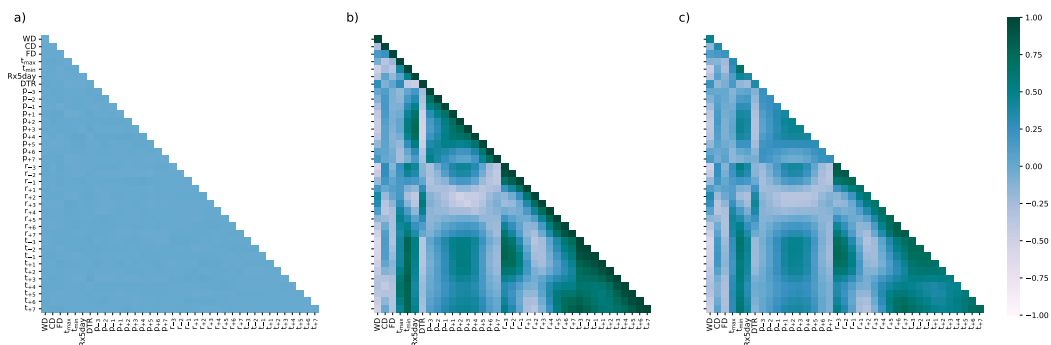
4

Figure S9: Correlation plots for (a): datapoints after permuting using random 20-fold CV, (b): the original data, (c): datapoints after permuting using feature-clusters CV.
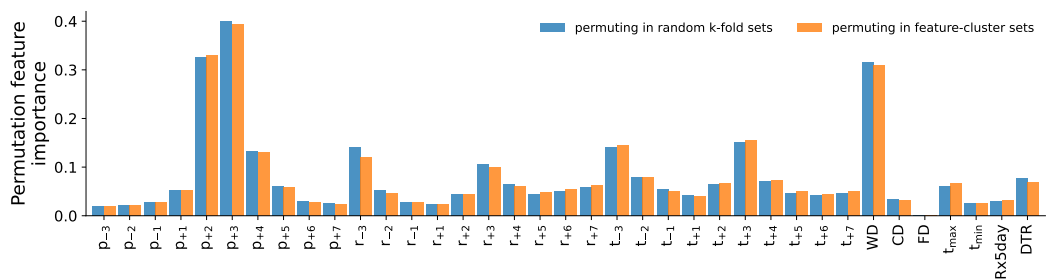

Figure S10: Permutation feature importances calculated using random 20-fold CV, with permuted datapoints generated using random 20-fold CV or using feature-clusters CV.