Supplement of Geosci. Model Dev., 17, 2663–2682, 2024
https://doi.org/10.5194/gmd-17-2663-2024-supplement

Geoscientific Model Development
Open Access

Supplement of

Carbon Monitor Power-Simulators (CMP-SIM v1.0) across countries: a data-driven approach to simulate daily power generation

Léna Gurriaran et al.

Correspondence to: Léna Gurriaran (lena.gurriaran@lsce.ipsl.fr, lena.gurriaran@atos.net)

Supplementary Materials

S1 Description of the hyperparameters optimized through the grid search process

S1.1 Random Forest

Random forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. The principle behind random forest is to create a diverse set of decision trees that are trained on different subsets of the training data and features.

At each node of each tree, a random subset of features is considered for splitting, which introduces randomness into the model and reduces the variance of the individual trees. The final prediction of the random forest is the majority vote of the predictions made by all the trees in the forest.

Max depth: It controls the complexity of the individual trees in the ensemble by limiting the number of levels in each tree, which in turn limit the number of splits a leaf can make.

Min Samples Leaf: It specifies the minimum number of samples that should be present in a leaf node of a tree. It controls the granularity of the splits made by the tree, and thus, it affects the complexity of the tree.

Min Samples Split: It specifies the minimum number of samples that should be present in an internal node before it can be split. It controls the overall structure of the tree and affects the degree of generalization of the tree.

N Estimator: It specifies the number of trees in the forest

S1.2 Gradient Boosting

Gradient boosting is an ensemble learning algorithm that builds a sequence of decision trees to make predictions. The principle behind gradient boosting is to sequentially add decision trees to the ensemble that corrects the errors made by the previous trees.

At each iteration, the algorithm fits a decision tree to the residuals or gradients of the previous model's predictions. The residuals represent the difference between the true labels and the current model's predictions, while the gradients represent the first-order derivatives of the loss function with respect to the model's predictions. By fitting trees to the residuals or gradients, the algorithm can focus on the examples that are difficult to predict and improve the overall performance of the model.

Max Depth: It controls the maximum depth of each decision tree in the ensemble. It specifies the maximum number of levels that a tree can have, and it can be used to control the complexity of the model and prevent overfitting.

Learning Rate: It controls the contribution of each tree in the ensemble to the final prediction. It specifies the amount by which the predictions of the new trees are scaled before being added to the ensemble.

Max Iter: It controls the maximum number of trees in the ensemble. It specifies the number of boosting iterations that the model will fit to the data.

Max Leaf Nodes: It specifies the maximum number of terminal nodes or leaves that a tree can have, and it can be used to control the complexity of the model and prevent overfitting.
Min Samples Leaf: It specifies the minimum number of samples that should be present in a leaf node of a tree. It controls the granularity of the splits made by the tree, and thus, it affects the complexity of the tree.
L2 Regularization: It is a technique used to prevent overfitting by adding a penalty term to the objective function. The L2 penalty term is proportional to the sum of the squares of the model's parameters, and it encourages the model to choose smaller values for the parameters.

S1.3 MARS

The Multivariate Adaptive Regression Splines (MARS) model is a flexible non-parametric regression technique that can capture complex nonlinear relationships between predictors and a response variable. MARS is a form of regression that constructs a piecewise linear model by breaking the predictor space into smaller subspaces and fitting a linear regression model to each subspace.
The MARS model builds upon the basic concept of linear regression by introducing nonlinear features and interactions between variables. It works by iteratively identifying breakpoints or knots in the predictor variables and fitting linear regression models to each segment between the breakpoints.
MARS is designed to handle both continuous and categorical variables and can automatically detect interactions between them. The model starts by creating simple linear models for each predictor and then combines them to form a more complex model. The model uses a forward selection approach to determine which variables to include in the model and where to place the breakpoints.

Max Degree: It is the maximum degree of terms generated by the forward pass.
Penalty: The MARS algorithm constructs a model by building basis functions that are combinations of simple functions such as linear, hinge, or threshold functions. Each basis function is a product of one or more simple functions, and the number of basis functions can quickly grow with the number of predictors and interactions considered. To avoid overfitting and improve the model's generalization performance, the MARS model uses a regularization penalty that penalizes the complexity of the model. The regularization penalty term is added to the objective function of the model, and it is typically a function of the sum of the absolute values of the coefficients or weights of the basis functions. The penalty encourages the MARS algorithm to choose simpler models with fewer basis functions and smaller coefficients, thereby avoiding overfitting.

S1.4 GAM

GAM stands for Generalized Additive Models. It is a statistical model that extends the linear model by allowing for non-linear relationships between the dependent variable and one or more independent variables. The principle of GAM is based on the idea that a complex relationship between the response variable and the predictor variables can be modeled as a sum of smooth functions of the predictors.

The model assumes that the response variable is a function of the predictor variables, which can be modeled using a combination of smooth functions. These smooth functions can be linear, non-linear, or a combination of both, and can be modeled using a variety of techniques, such as cubic splines or smoothing splines.
The key principle of GAM is to use these smooth functions to capture the non-linear relationship between the dependent and independent variables, without imposing any specific functional form on the relationship. GAM models can be used for both regression and classification problems, and they are particularly useful for analyzing complex relationships that cannot be modeled using linear models.

Lambda: lambda refers to the smoothing parameter that controls the amount of smoothing applied to the smooth functions. Lambda is the parameter that determines the amount of smoothing applied to the smooth functions. A small value of lambda will result in less smoothing and a more complex, wiggly fit to the data, whereas a large value of lambda will result in more smoothing and a simpler, smoother fit to the data.

N Splines: Splines are flexible functions that can approximate a wide range of non-linear relationships between variables. N splines refer to the number of spline basis functions used to model a smooth function in GAM. A spline basis function is a mathematical function that defines the shape of the spline. A spline function is a linear combination of these basis functions.

S2. Outputs of the models for all the countries considered in this study.

EU27 & UK: ALE plots not shown in the main text

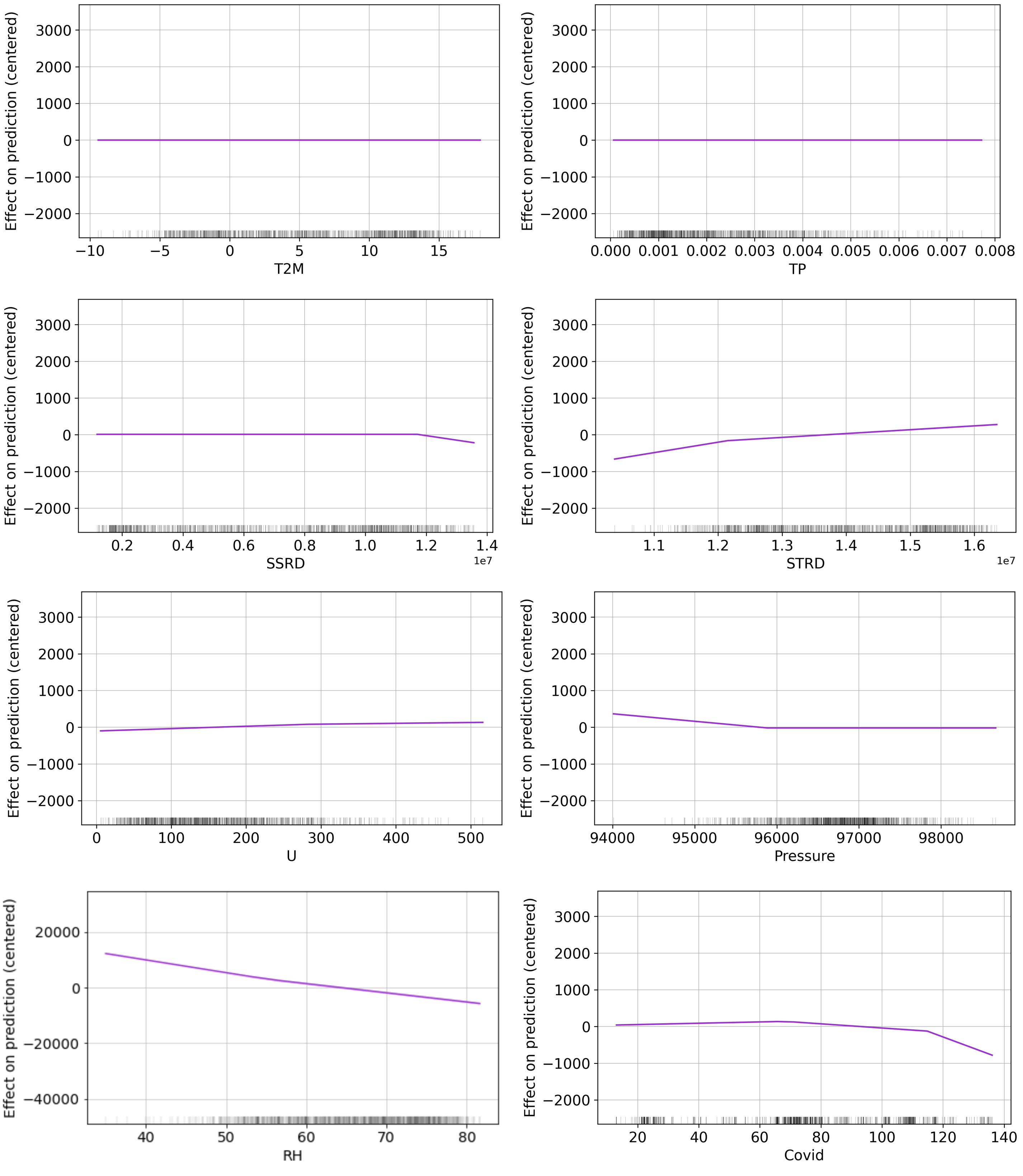


Figure S1. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

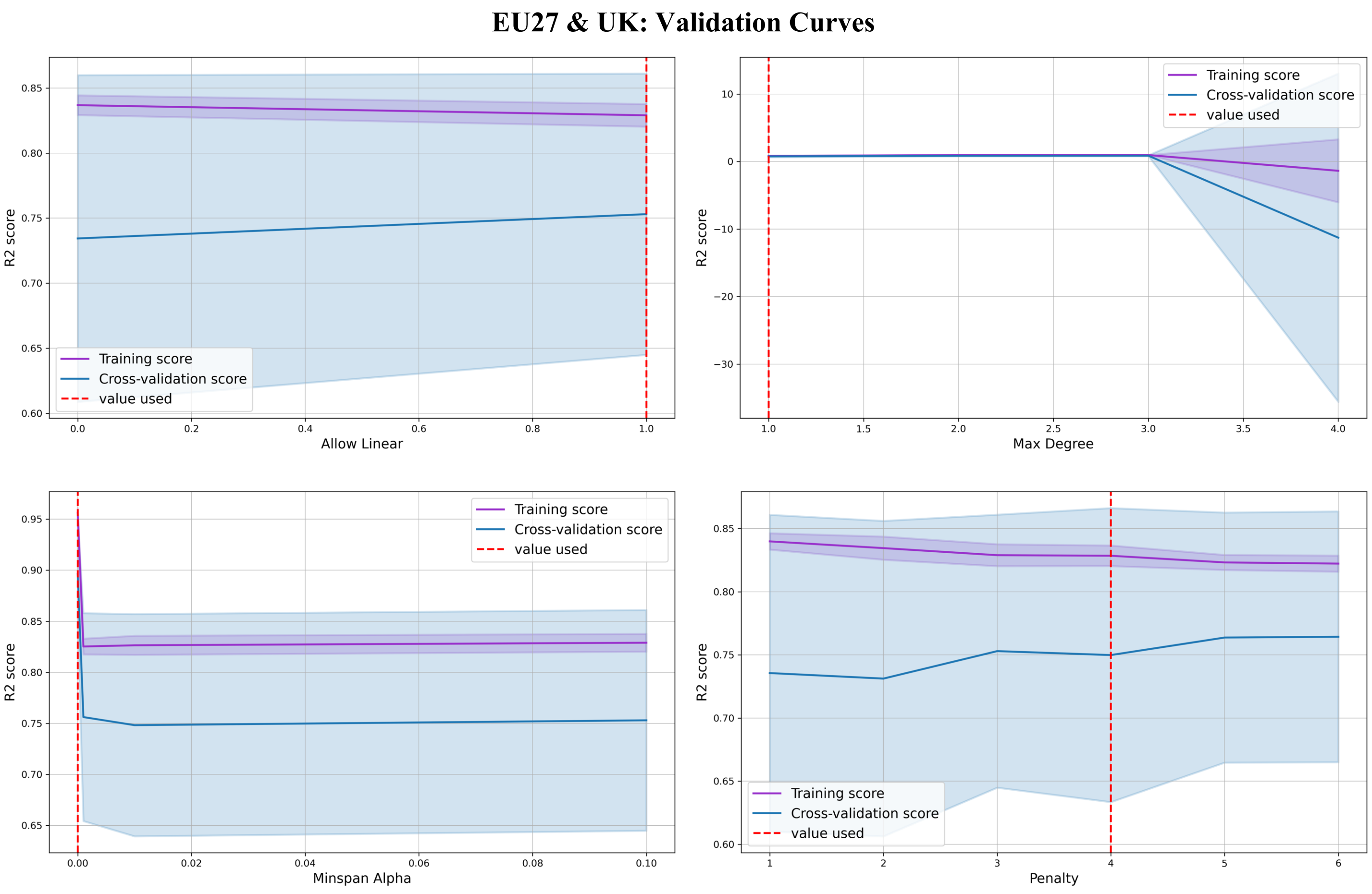


Figure S2. This figure displays validation curves for the hyperparameter of the MARS, which is the best model EU27 & UK. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

Australia: Input data

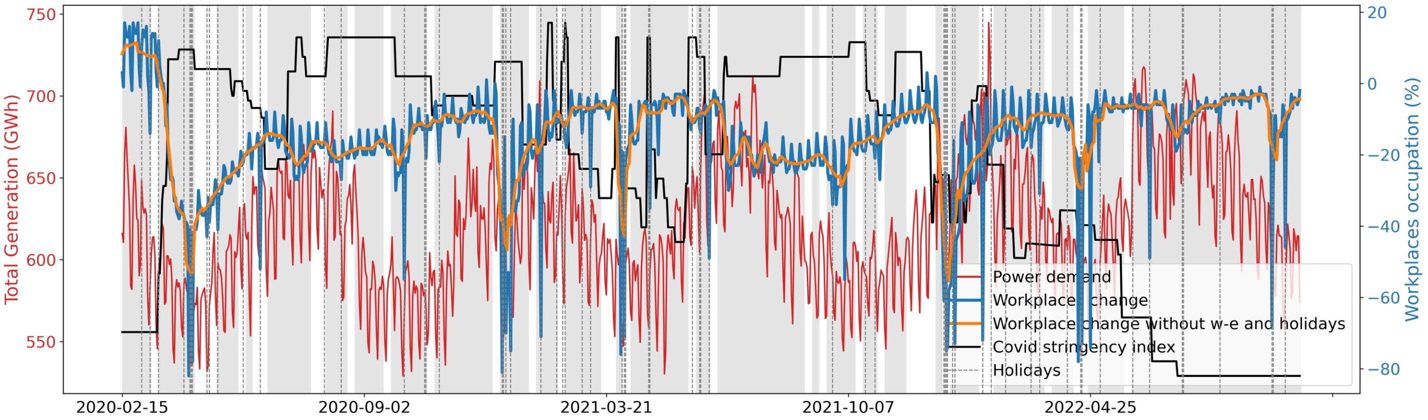


Figure S3. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

Australia: models performance

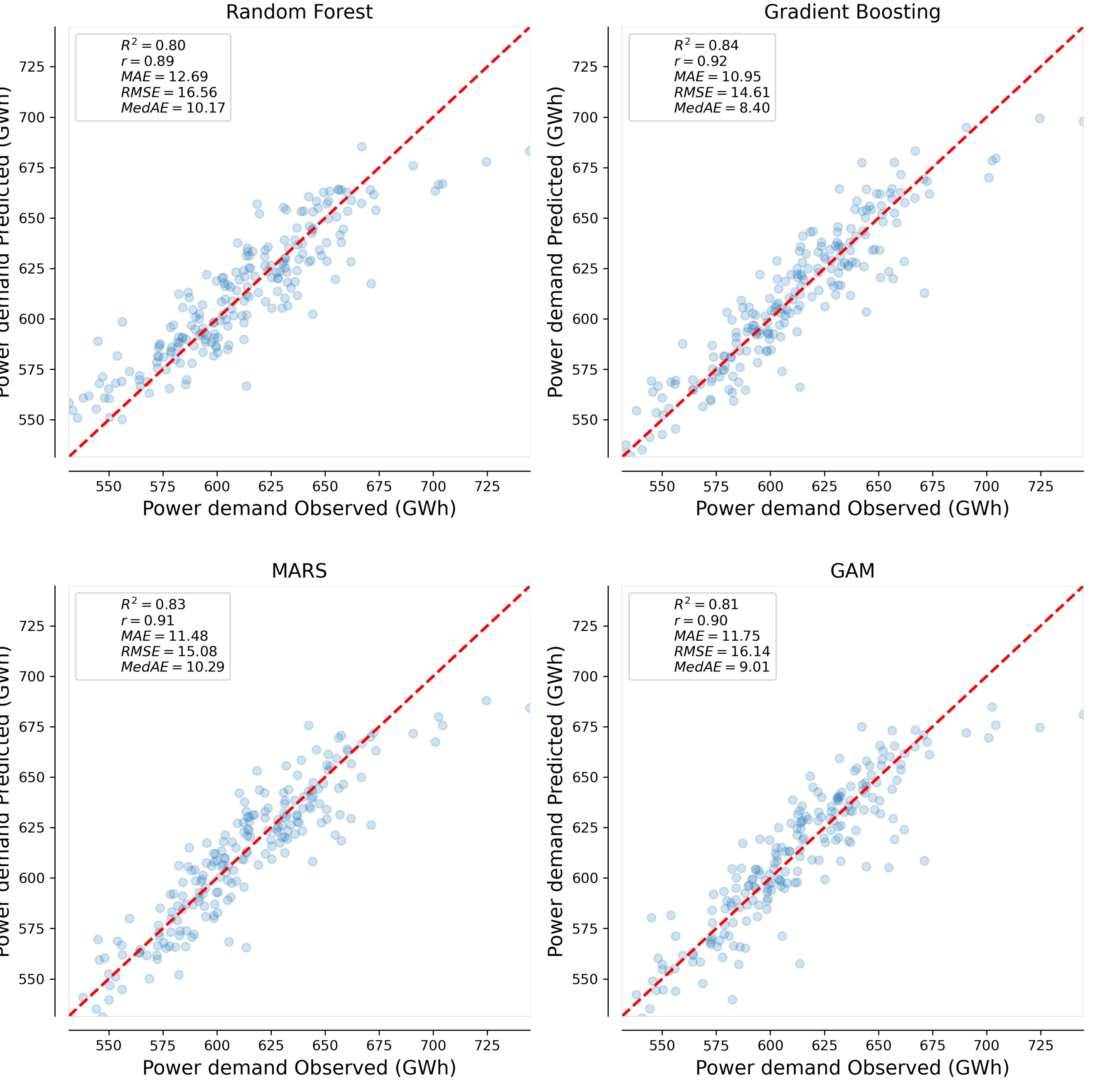


Figure S4. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

Australia: feature permutation importance

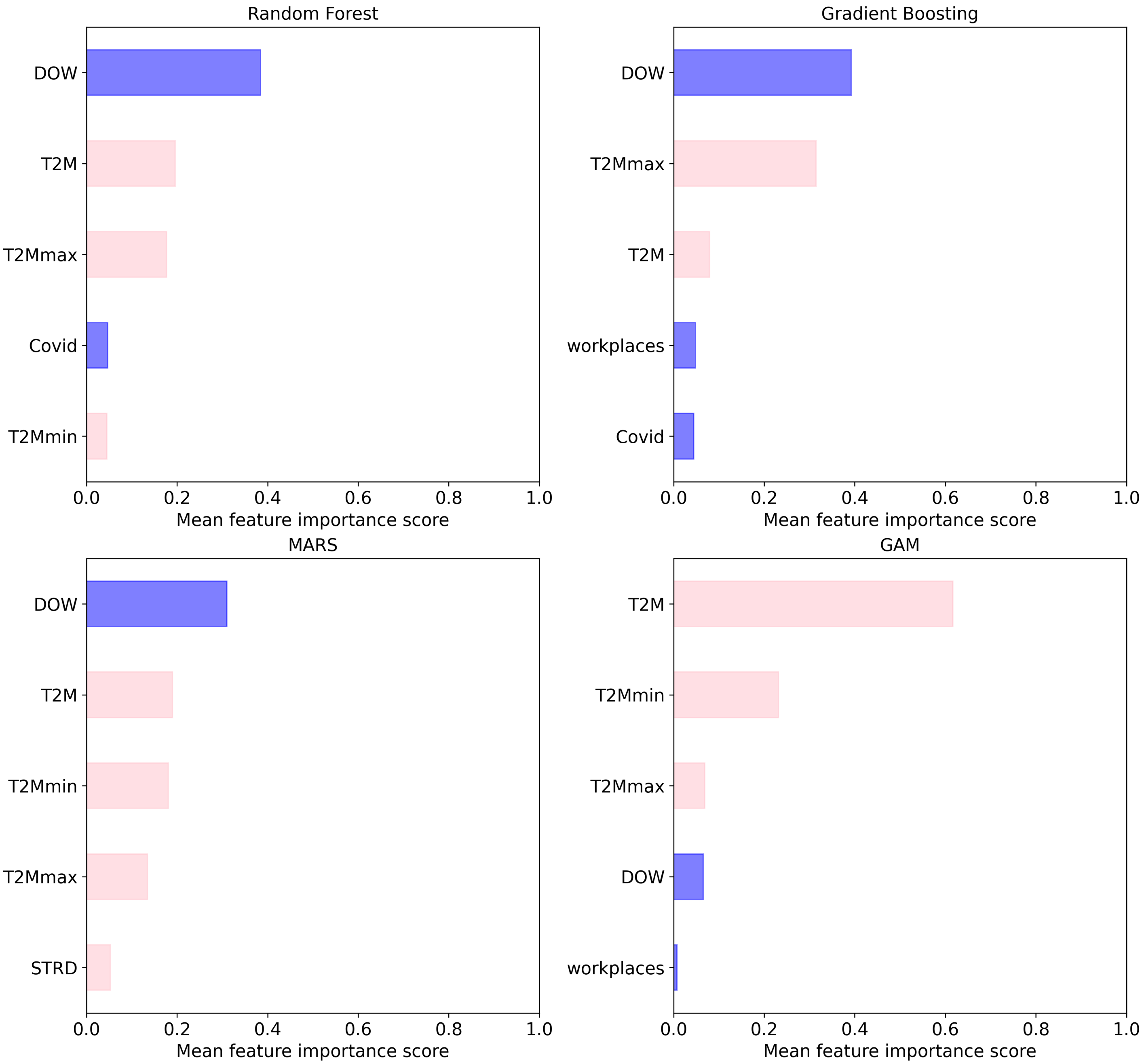
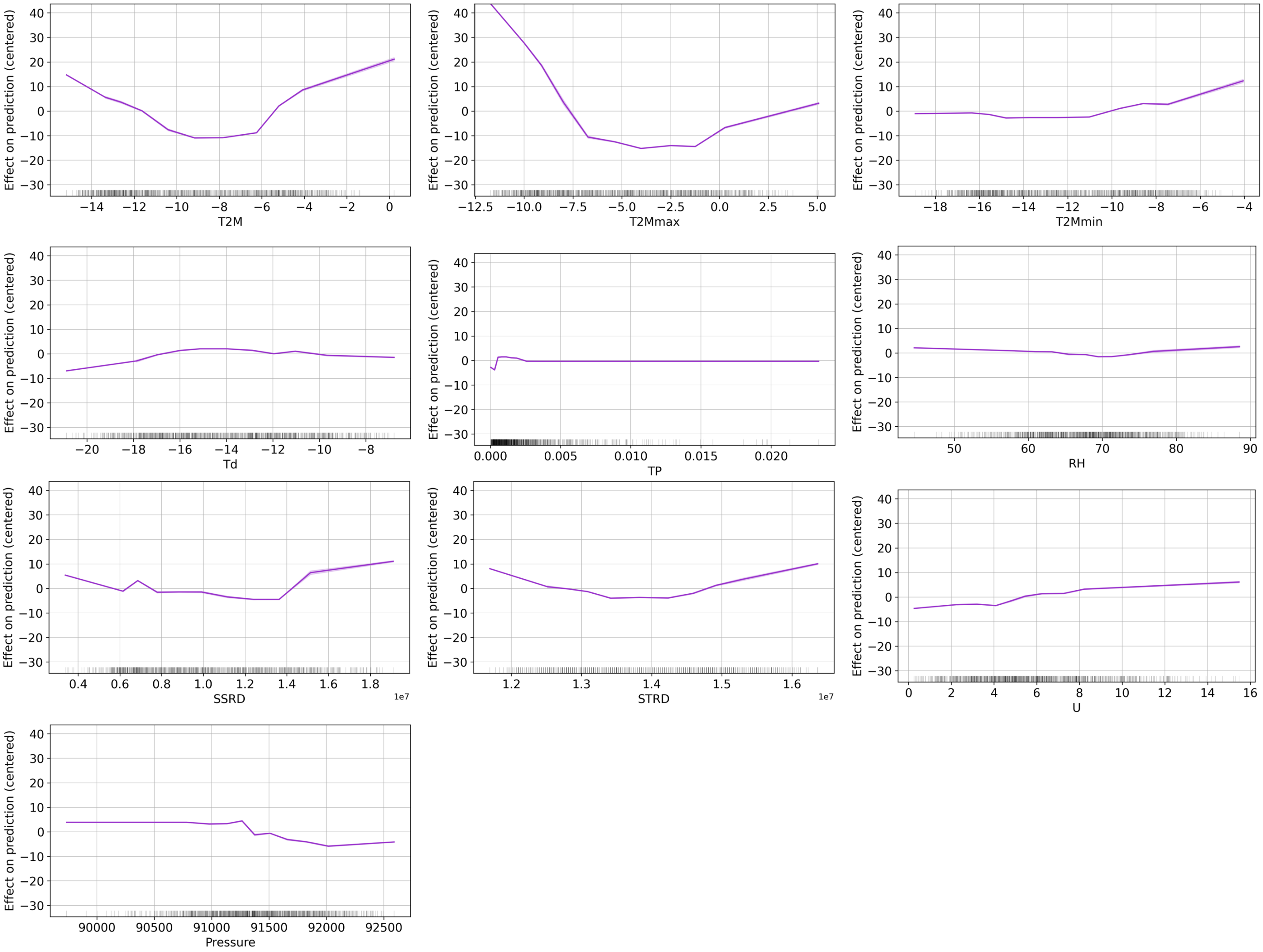


Figure S5. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

Australia: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

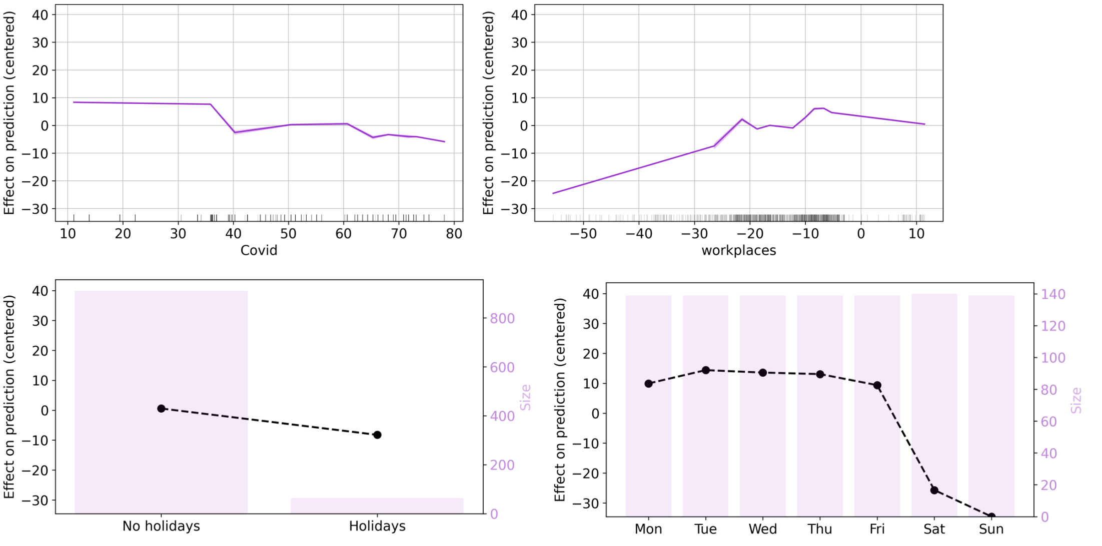


Figure S6. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

Australia: Validation curves

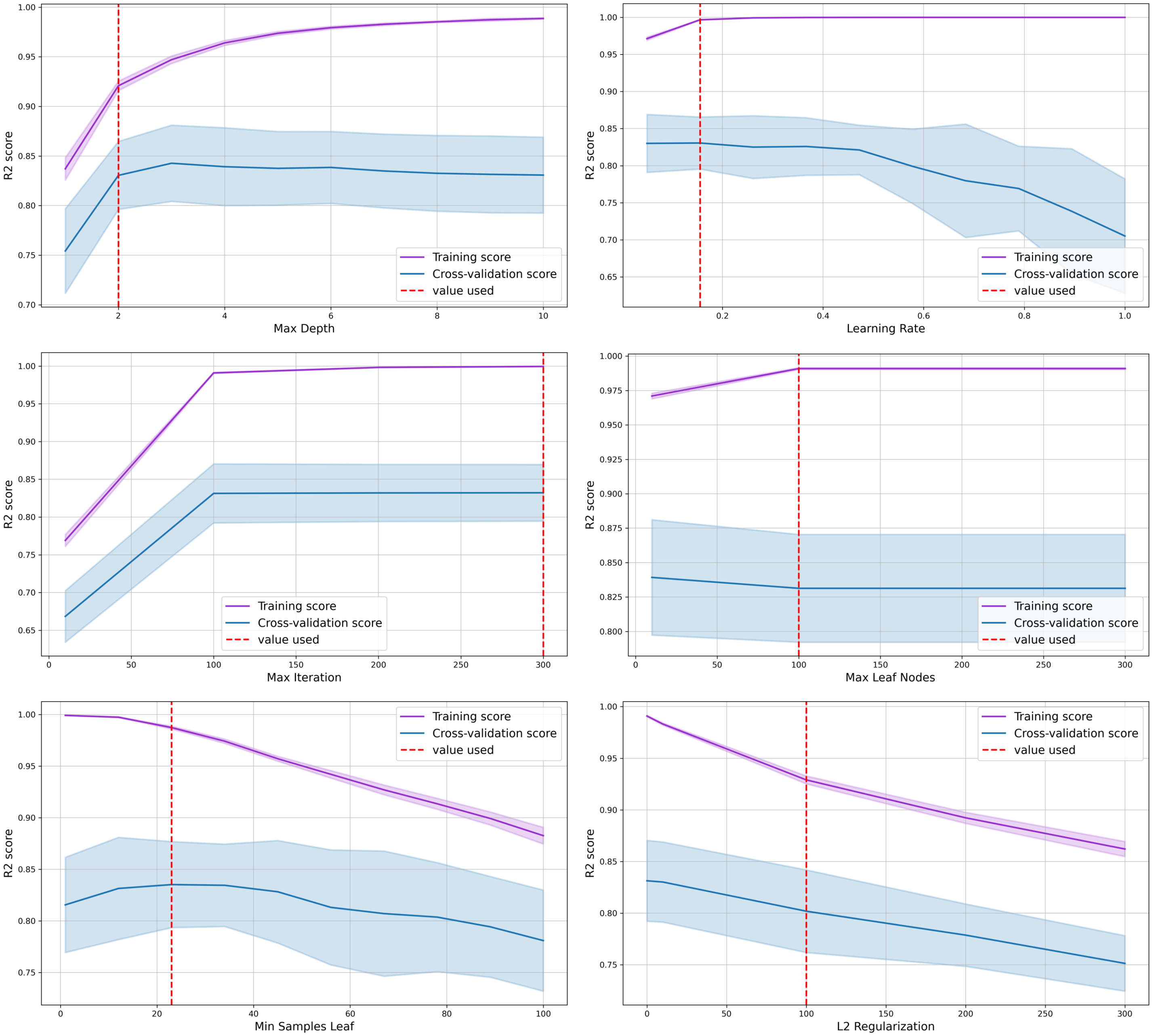


Figure S7. This figure displays validation curves for the hyperparameter of the Gradient Boosting, which is the best model in Australia. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

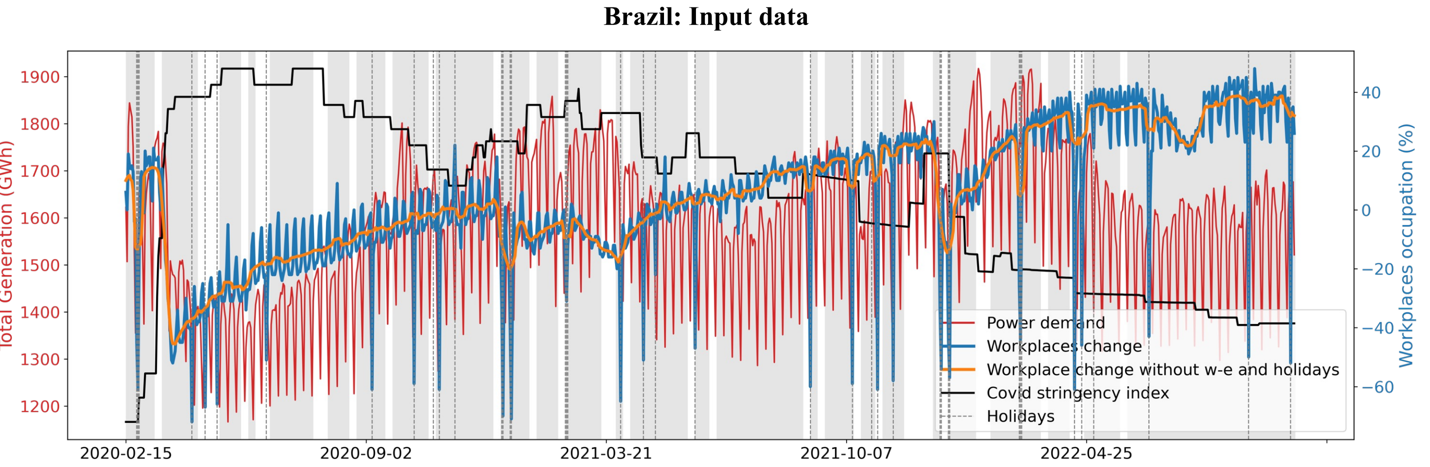


Figure S8. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

Brazil: models performance

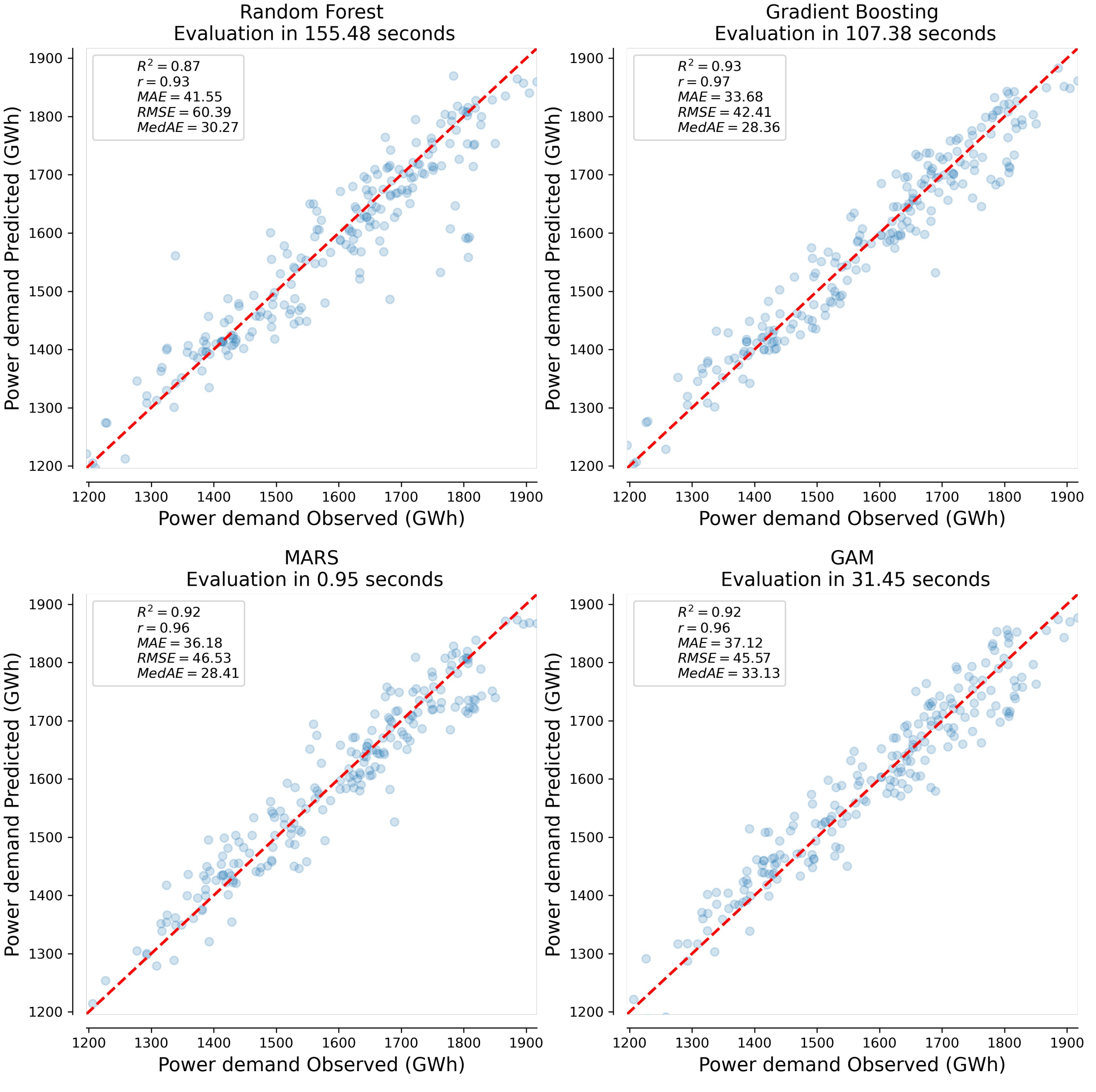


Figure S9. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

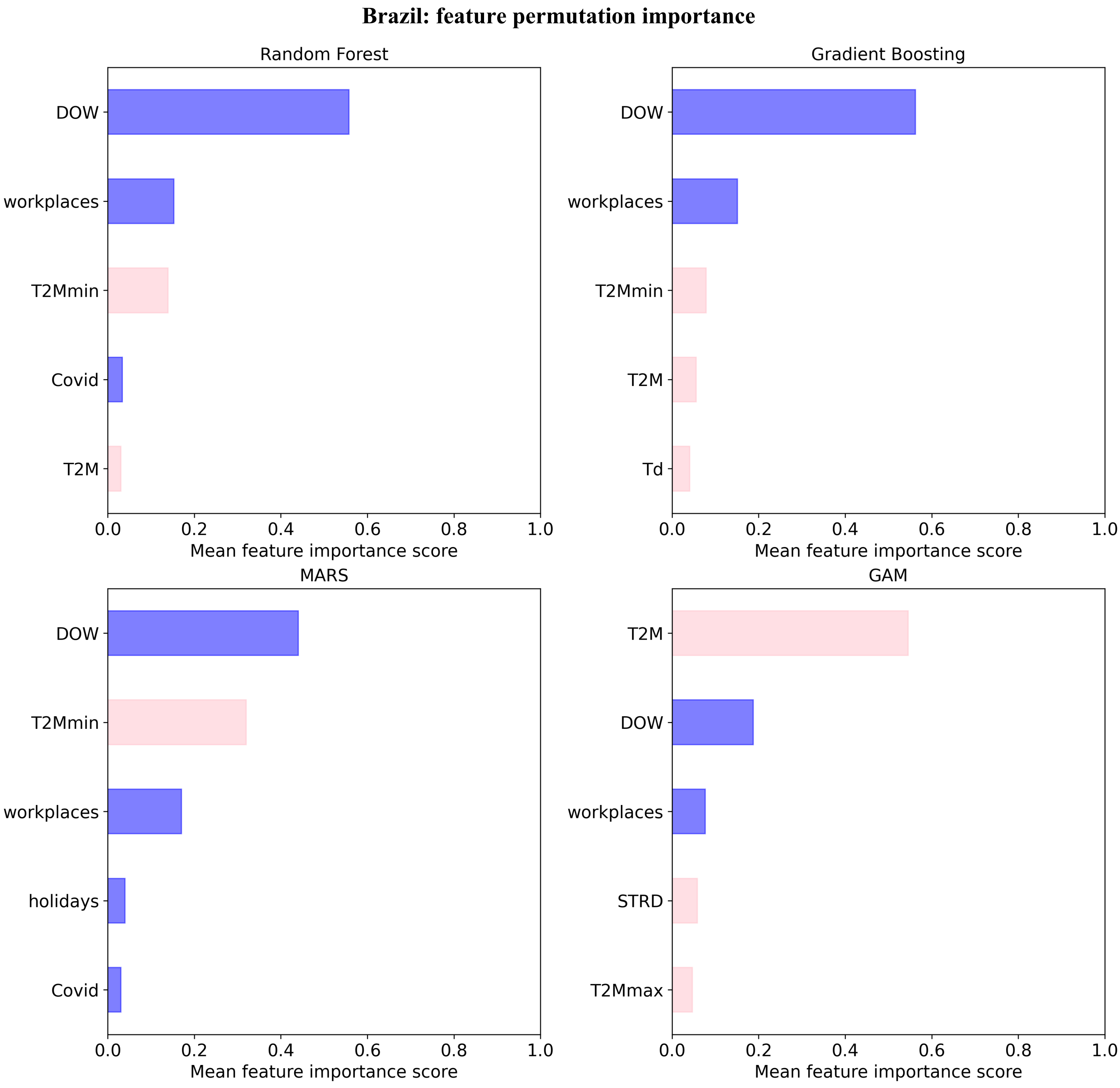
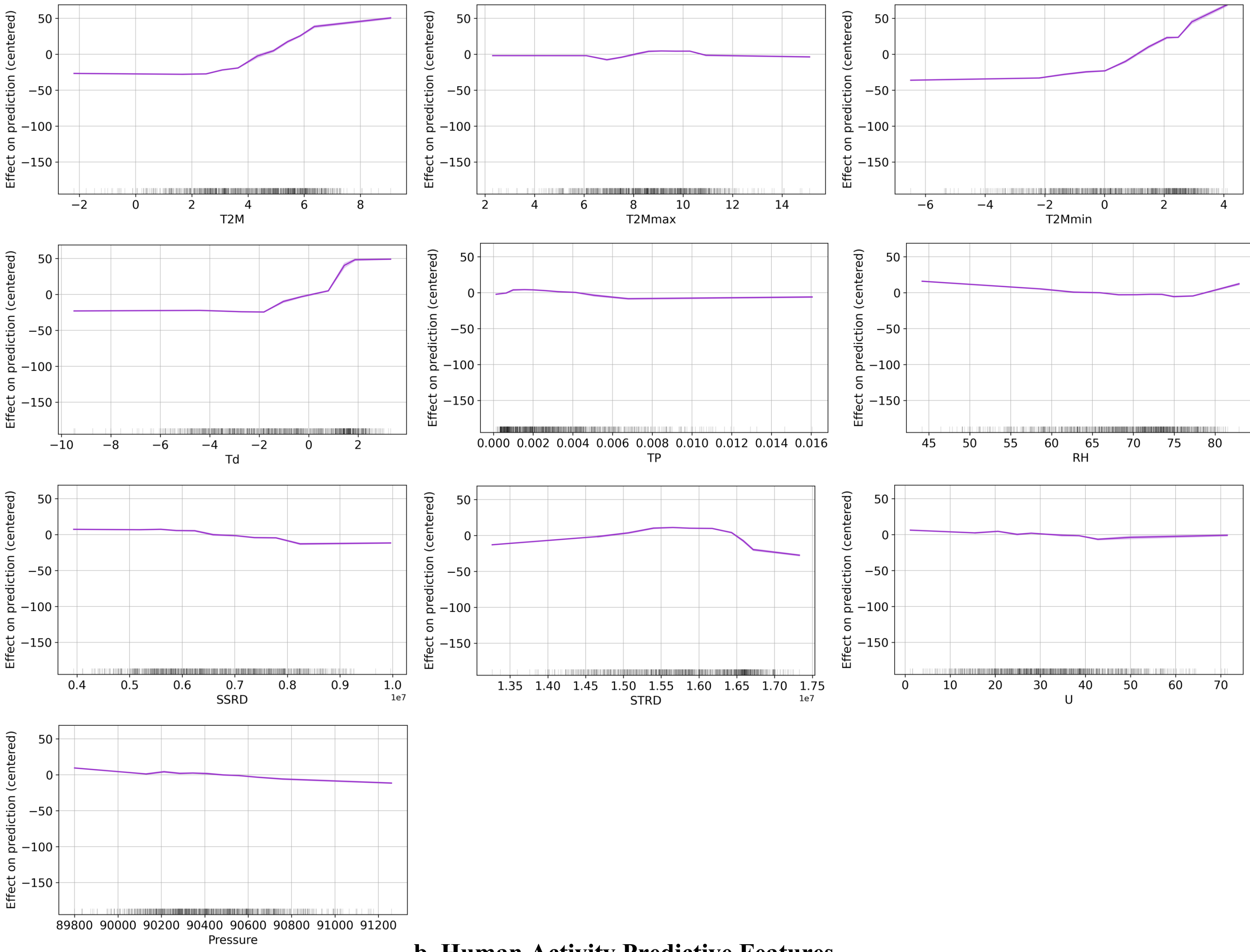


Figure S10.Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

Brazil: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

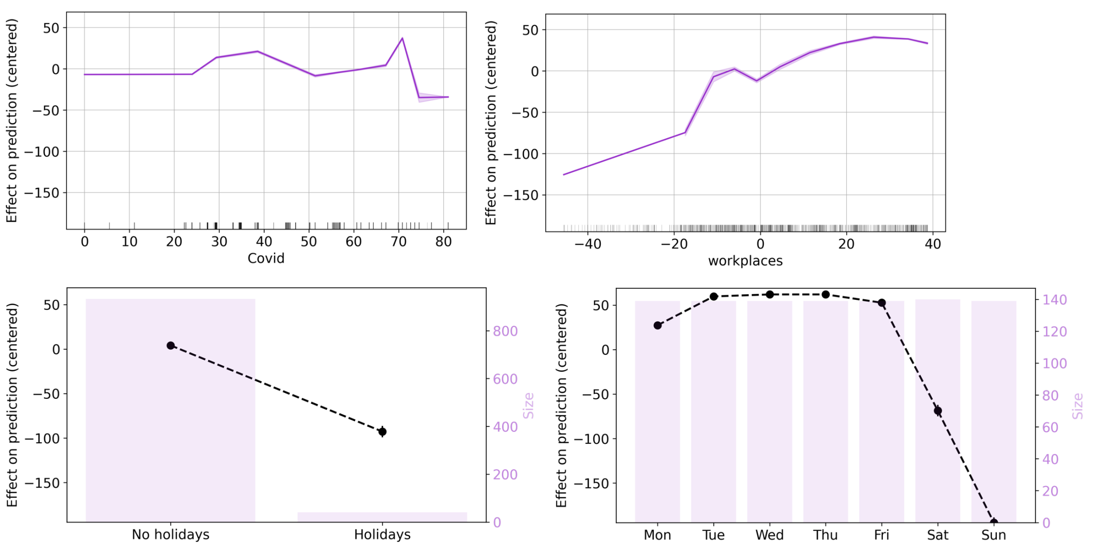


Figure S11. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

Brazil: Validation curves

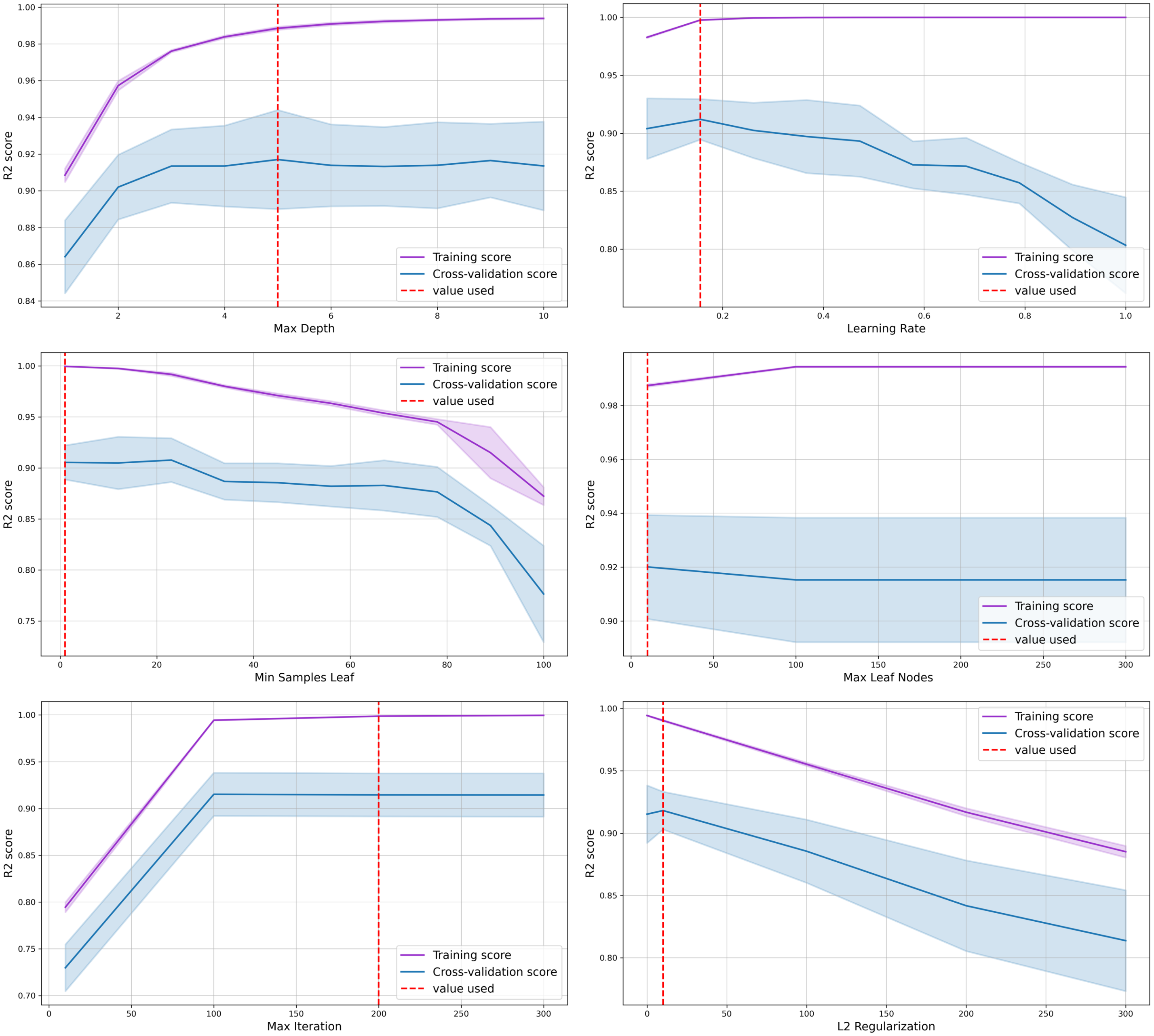


Figure S12. This figure displays validation curves for the hyperparameter of the gradient boosting model, which is the best model in Brazil. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

China: Input data

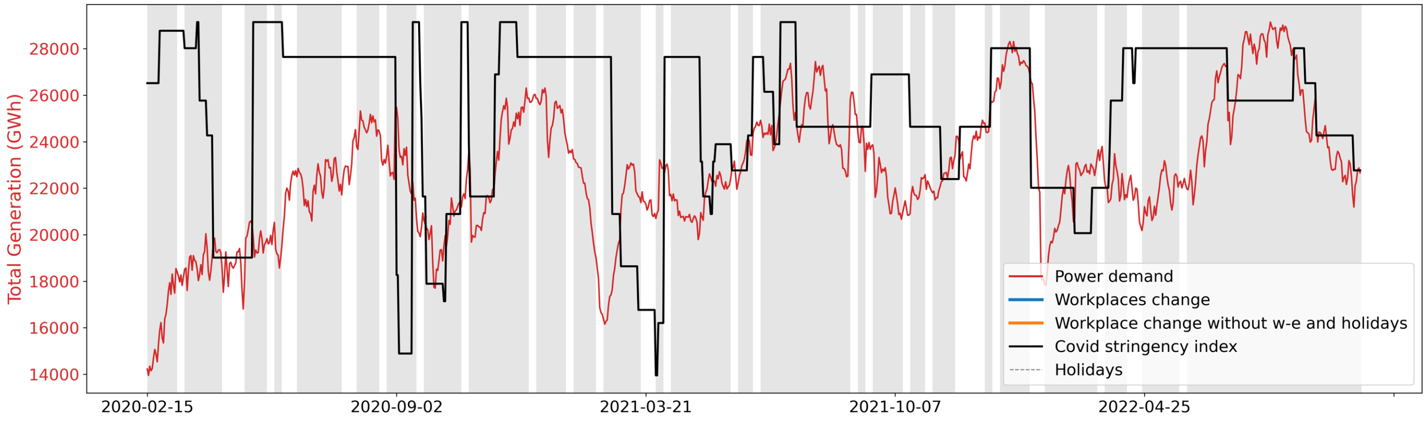


Figure S13. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

China: models performance

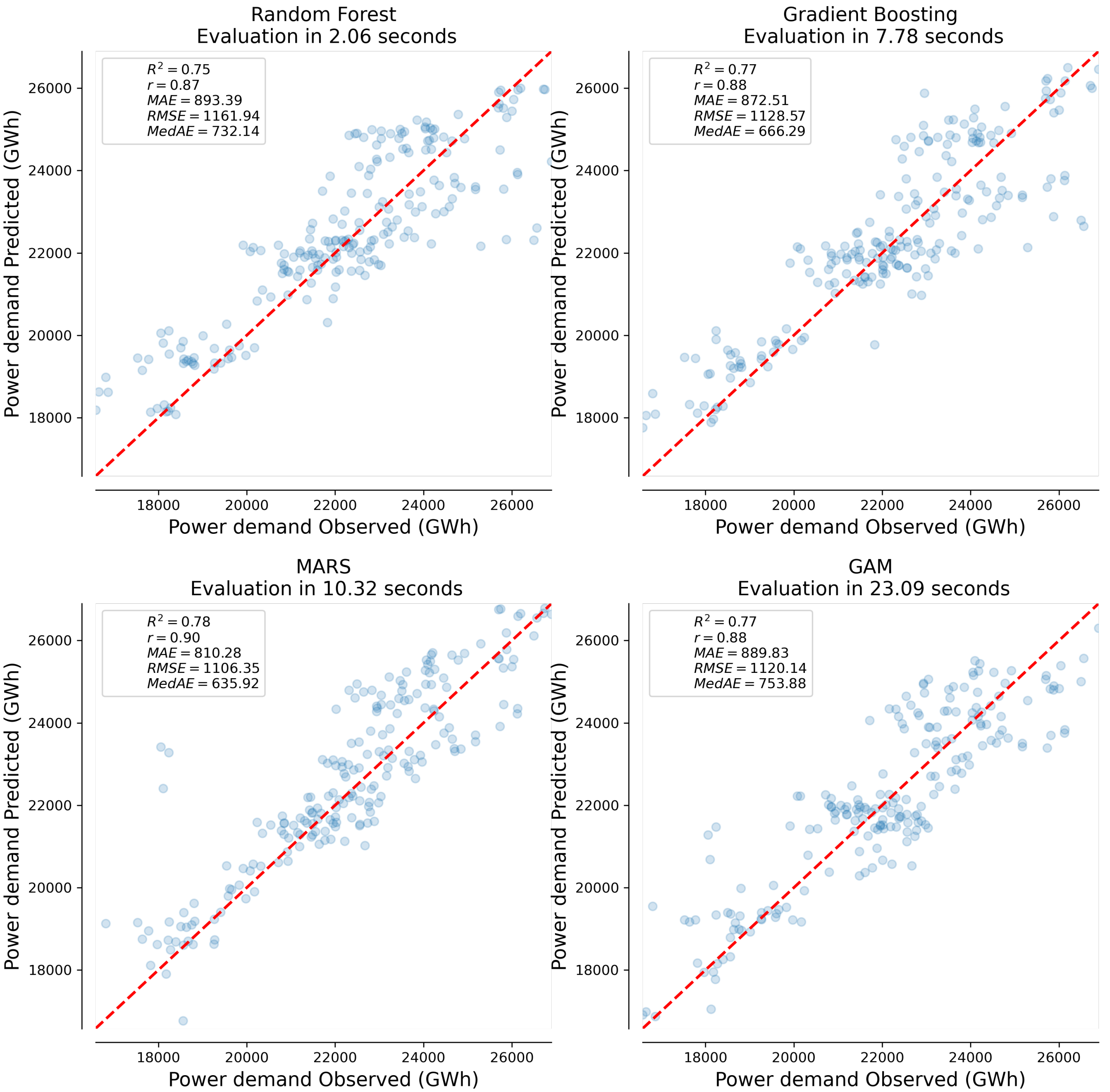


Figure S14. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

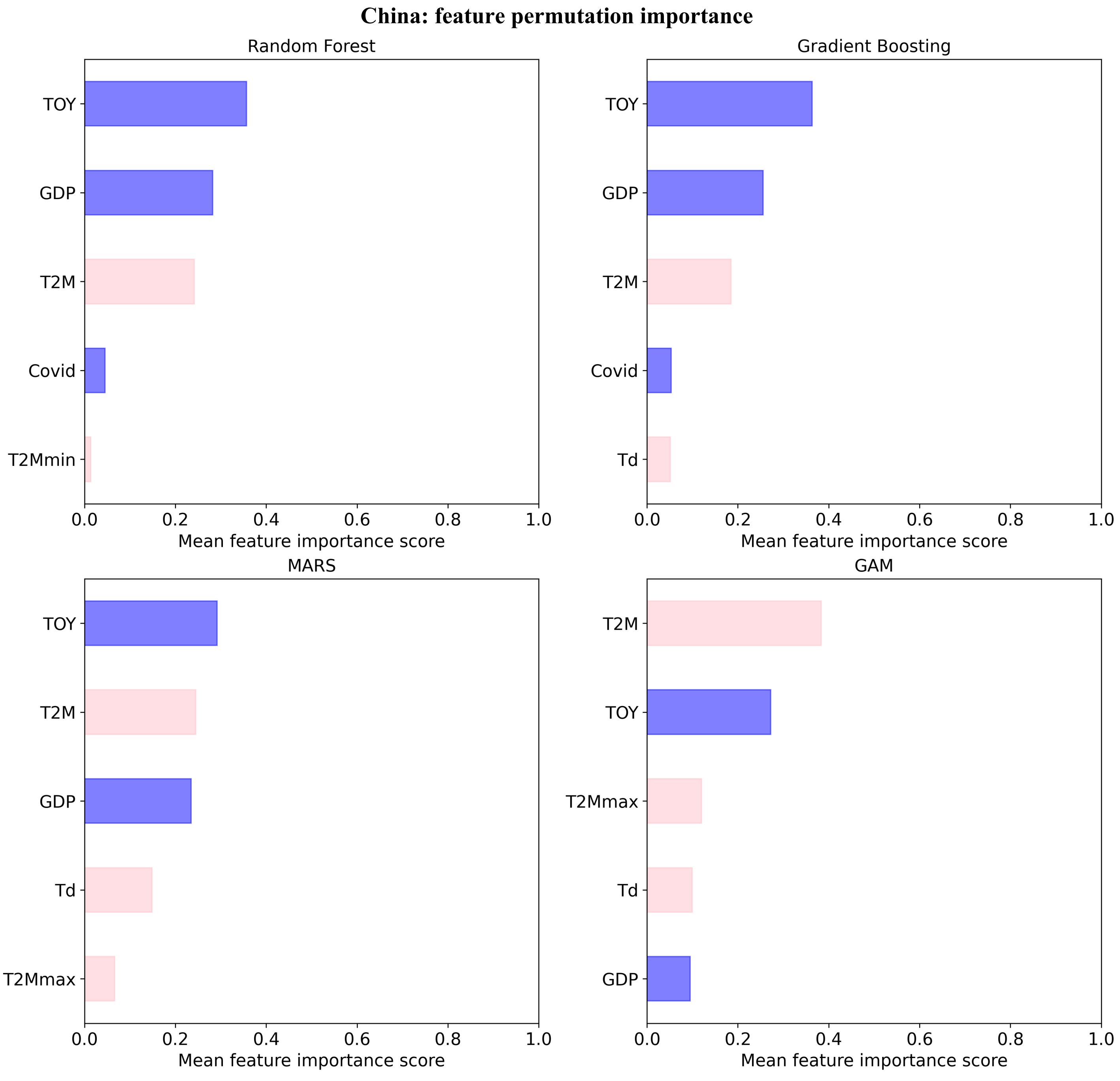
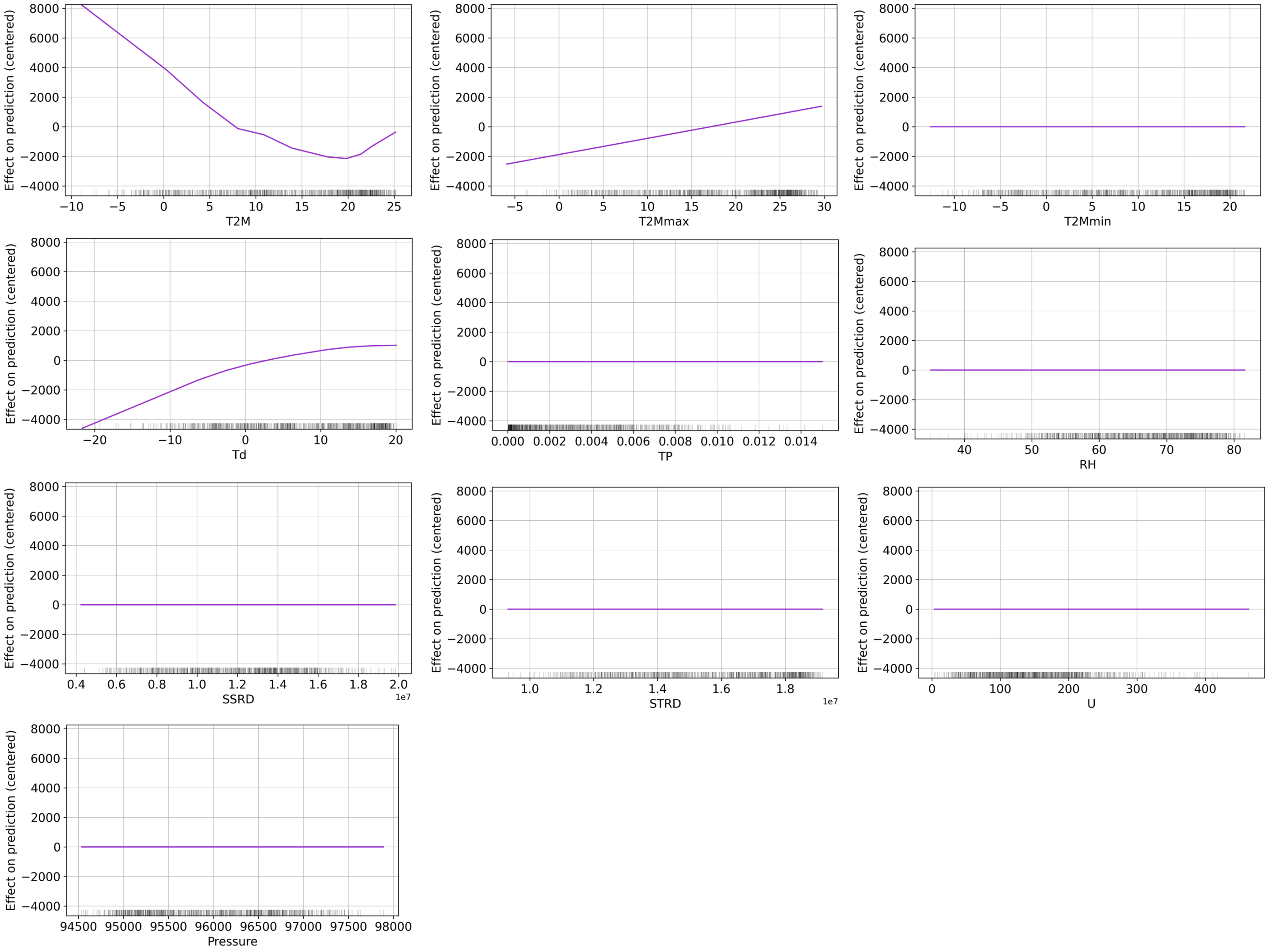


Figure S15. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

China: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

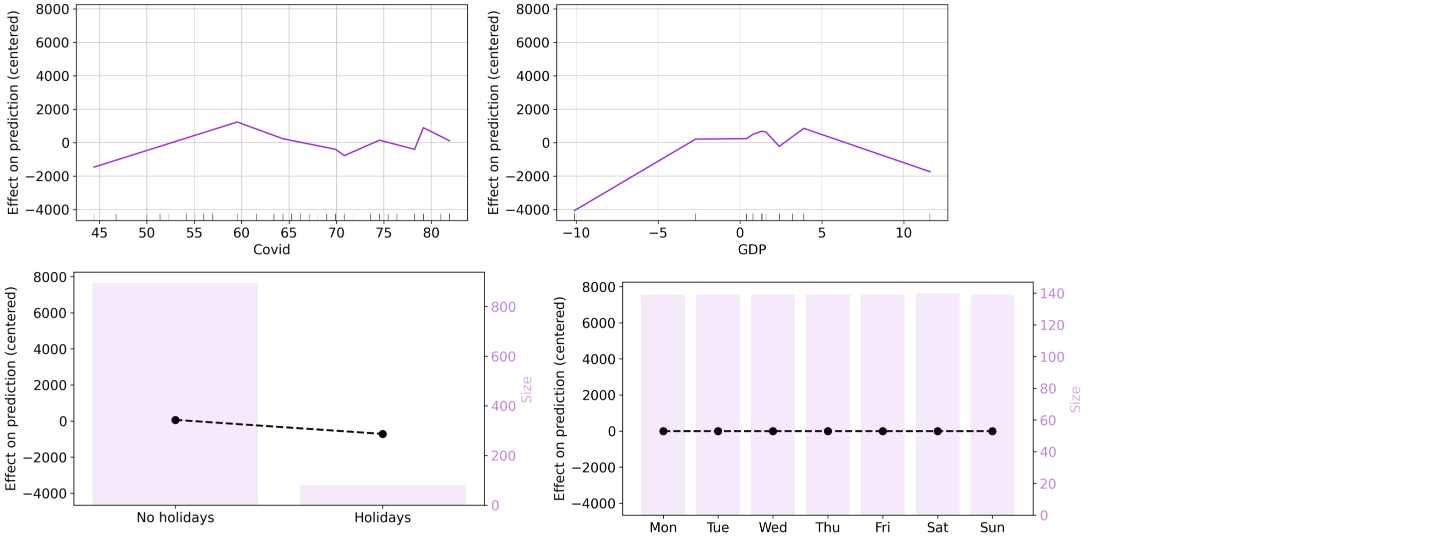


Figure S16. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

China: Validation curves

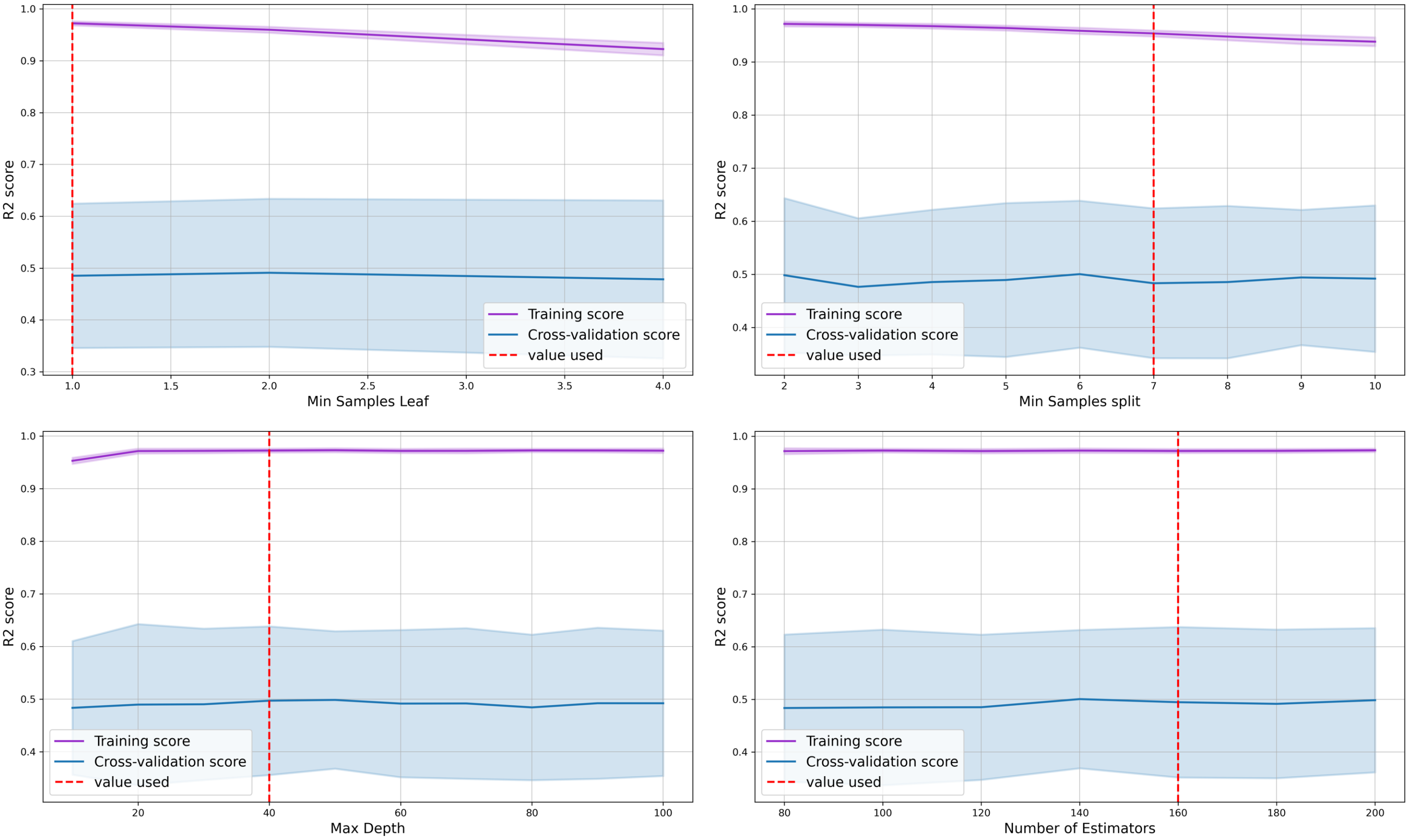


Figure S17. This figure displays validation curves for the hyperparameter of the MARS, which is the best model in China. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

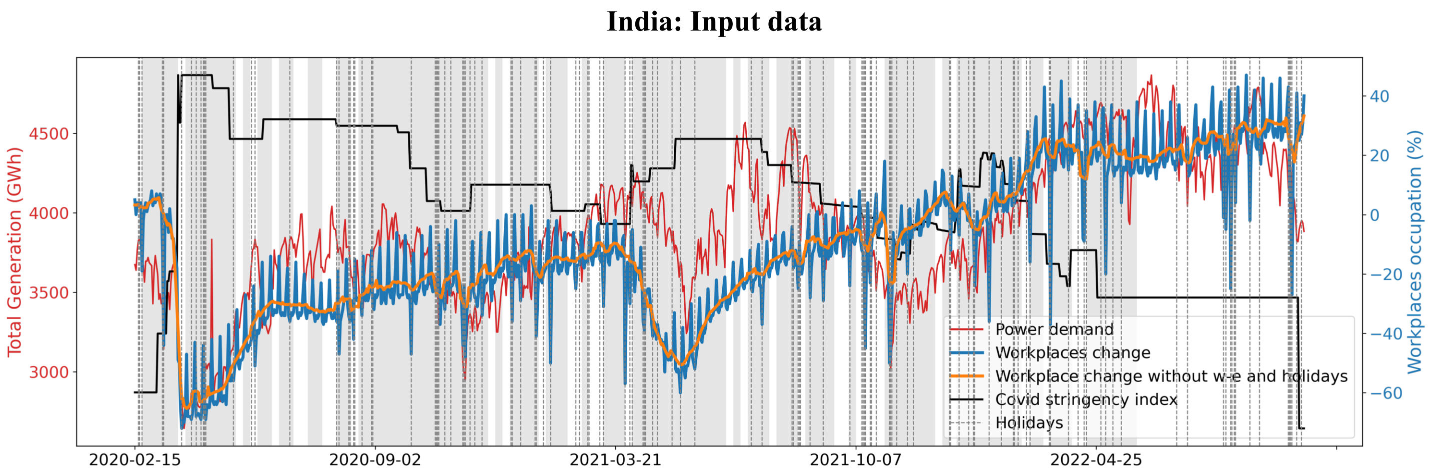


Figure S18. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

India: models performance

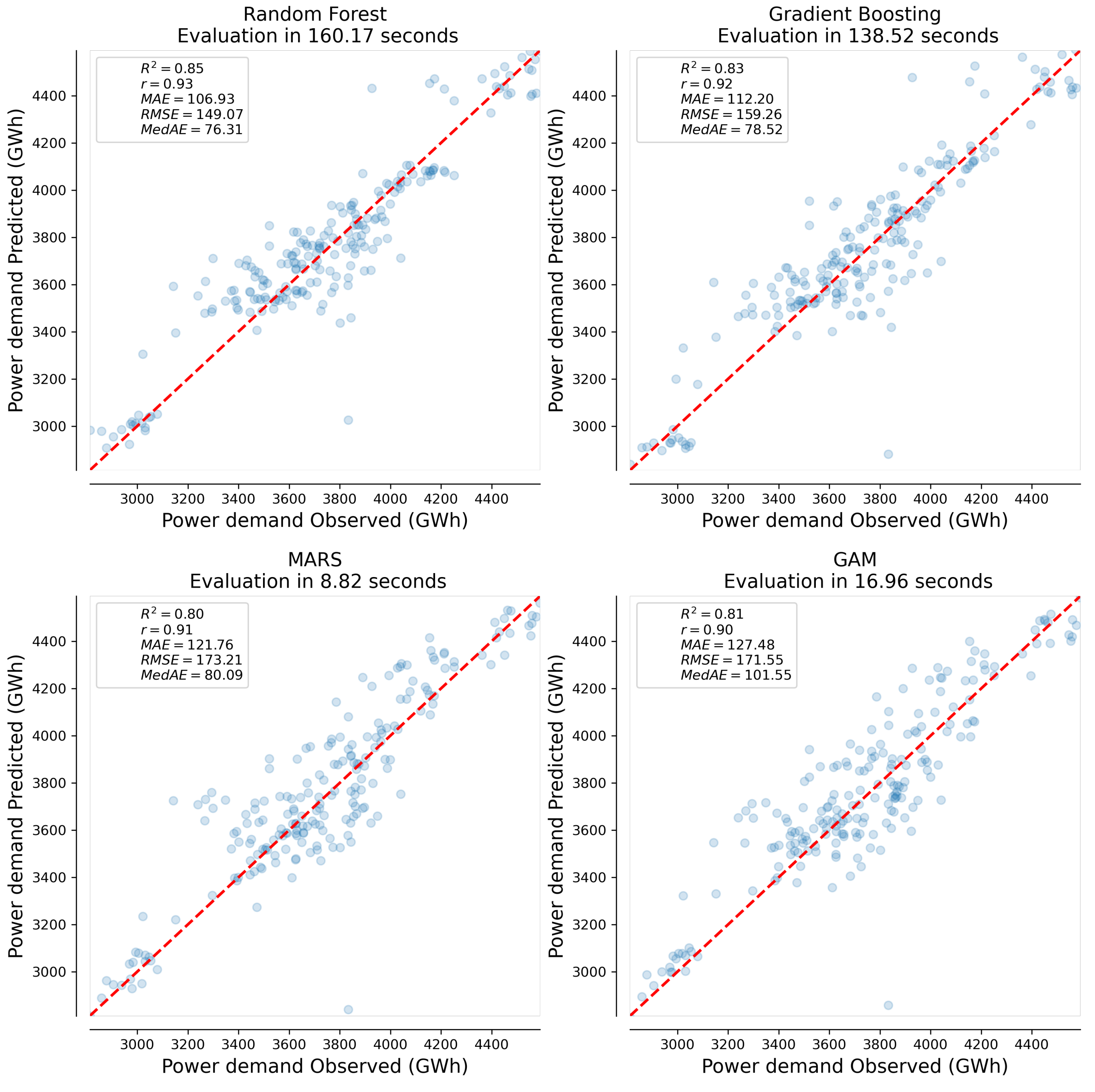


Figure S19. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

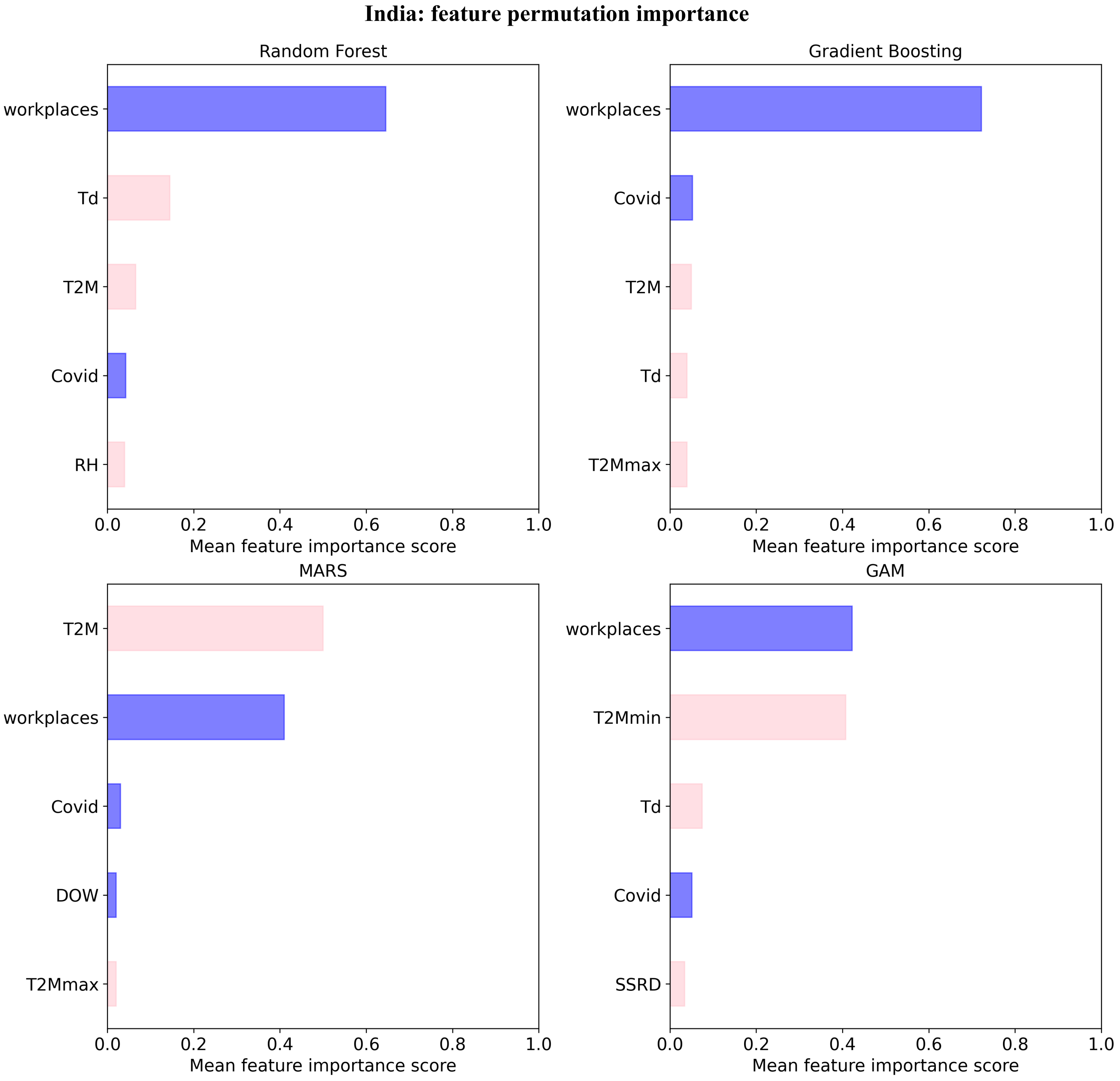
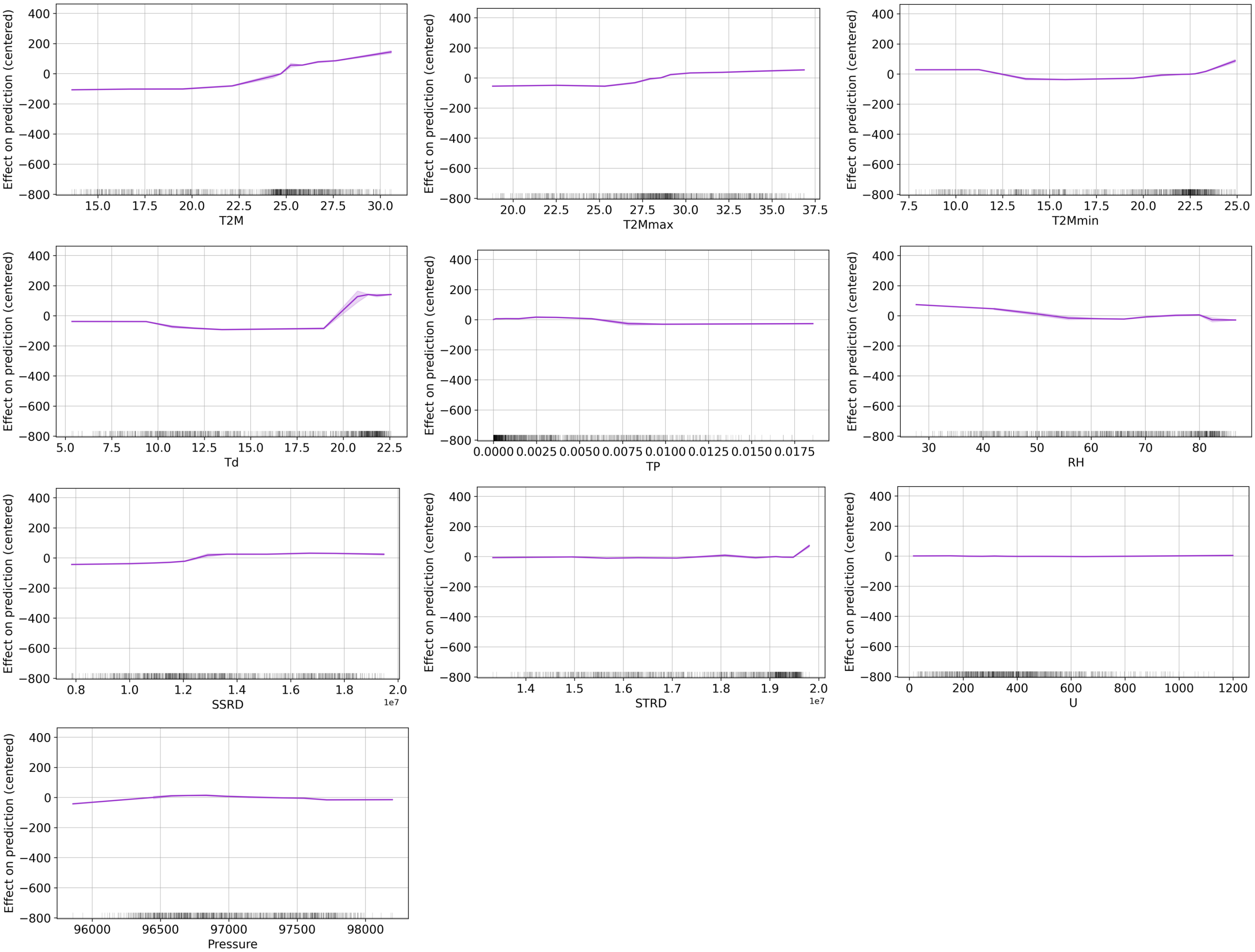


Figure S20. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

India: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

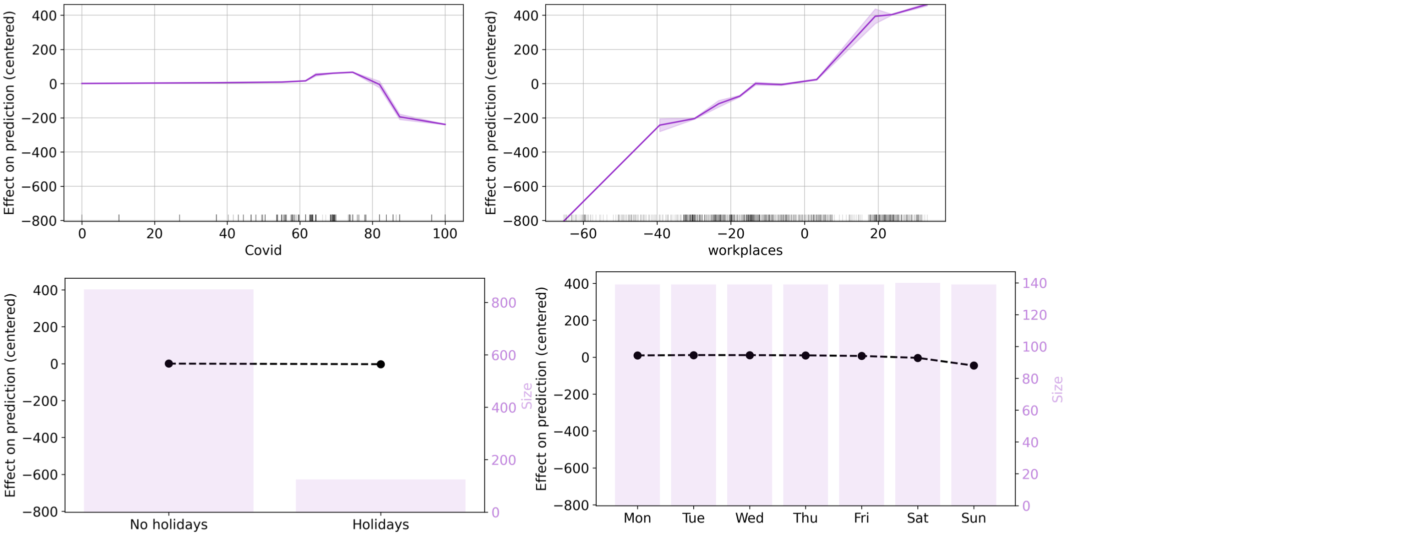


Figure S21. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

India: Validation curves

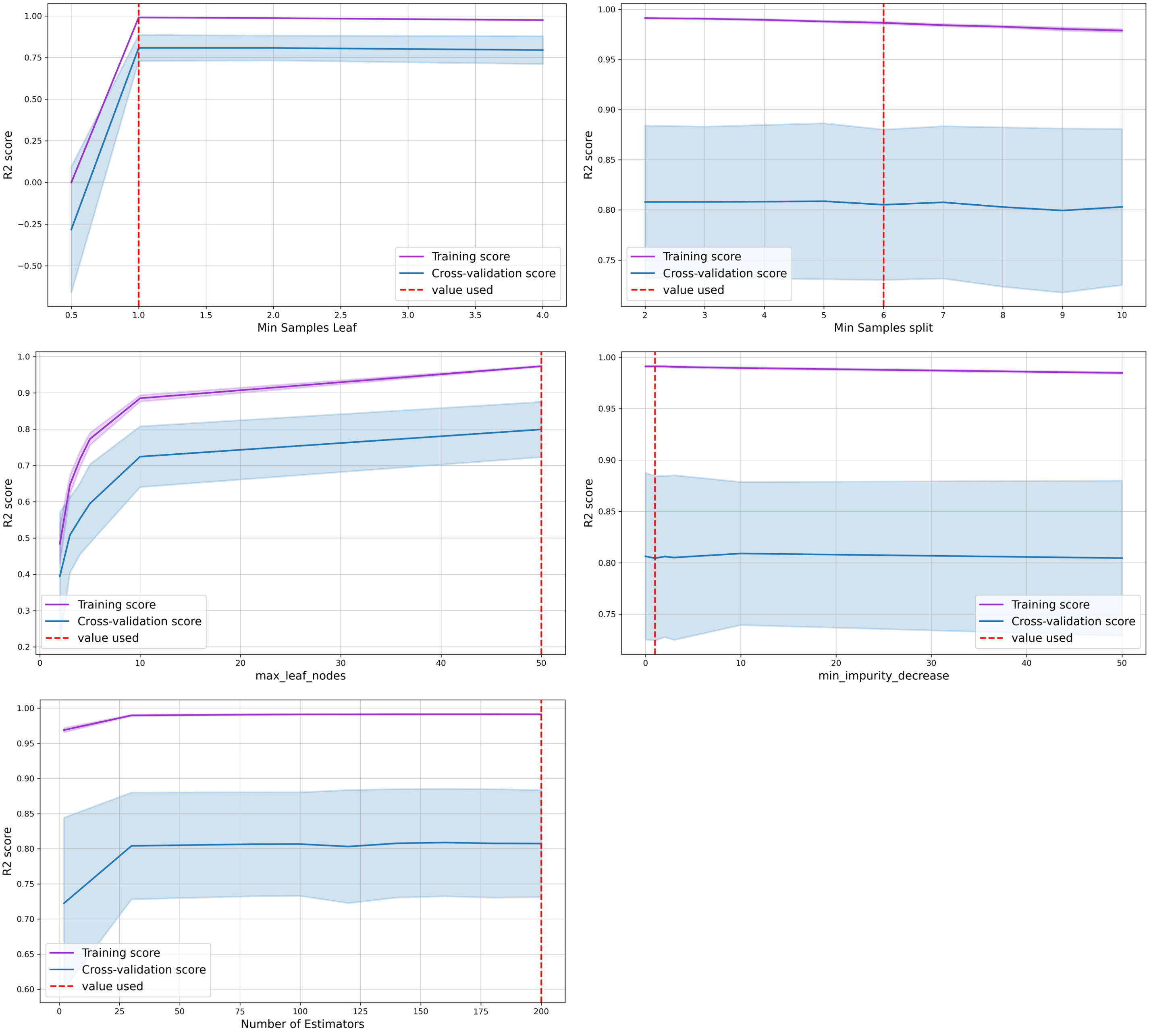


Figure S22. This figure displays validation curves for the hyperparameter of the Random Forest, which is the best model in India. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

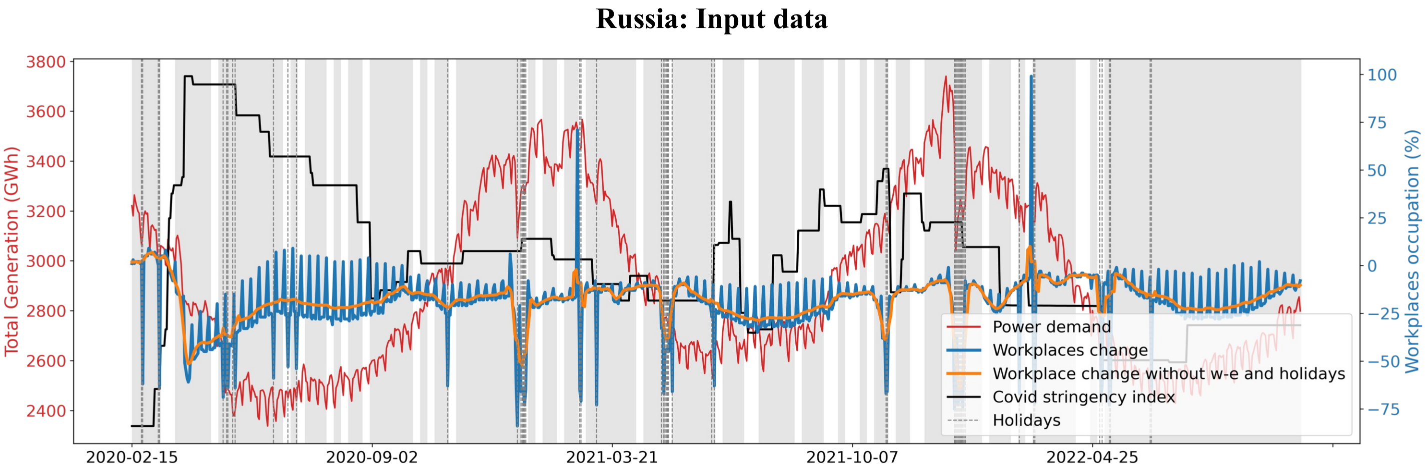


Figure S23. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

Russia: models performance

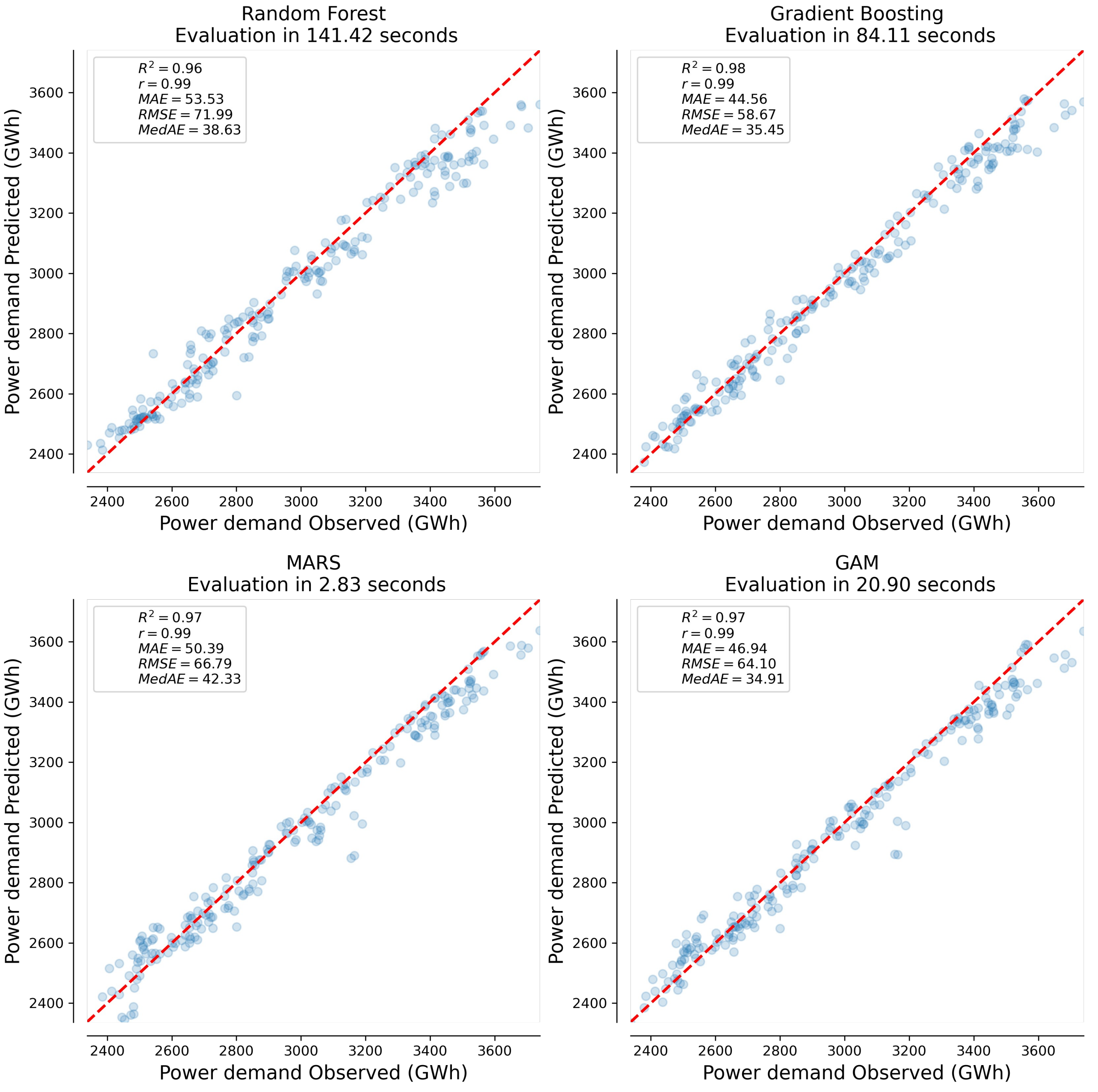


Figure S24. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

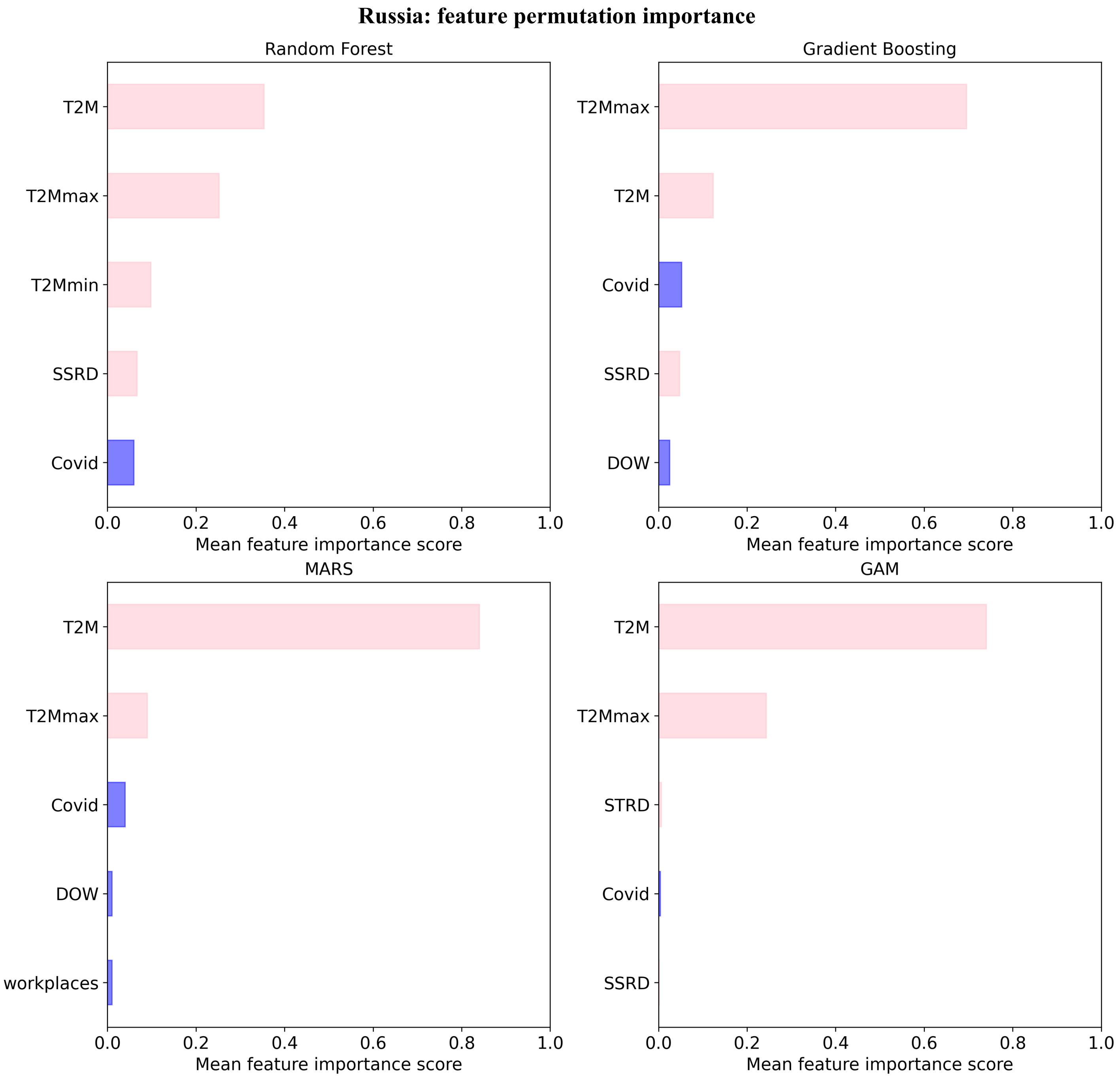
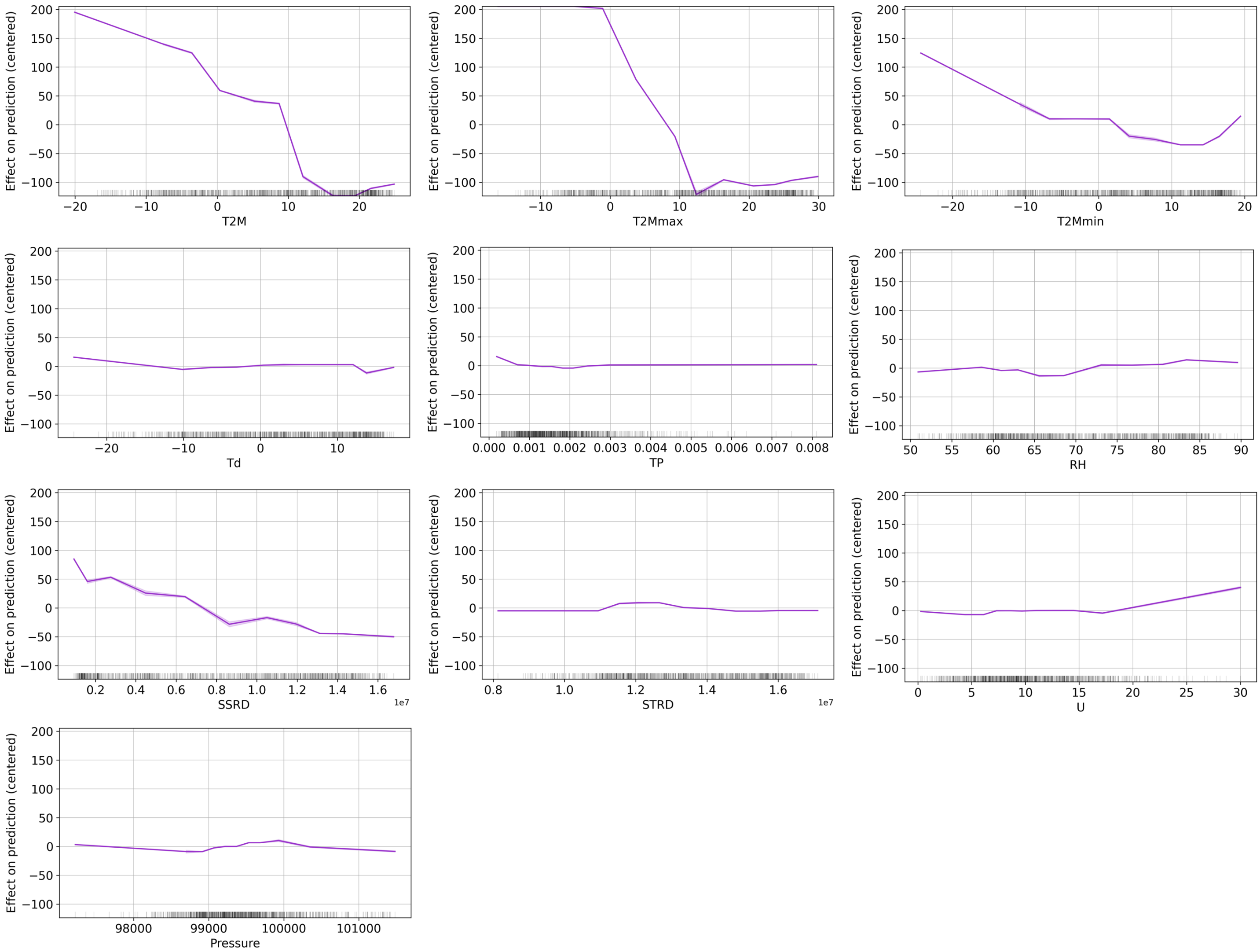


Figure S25. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

Russia: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

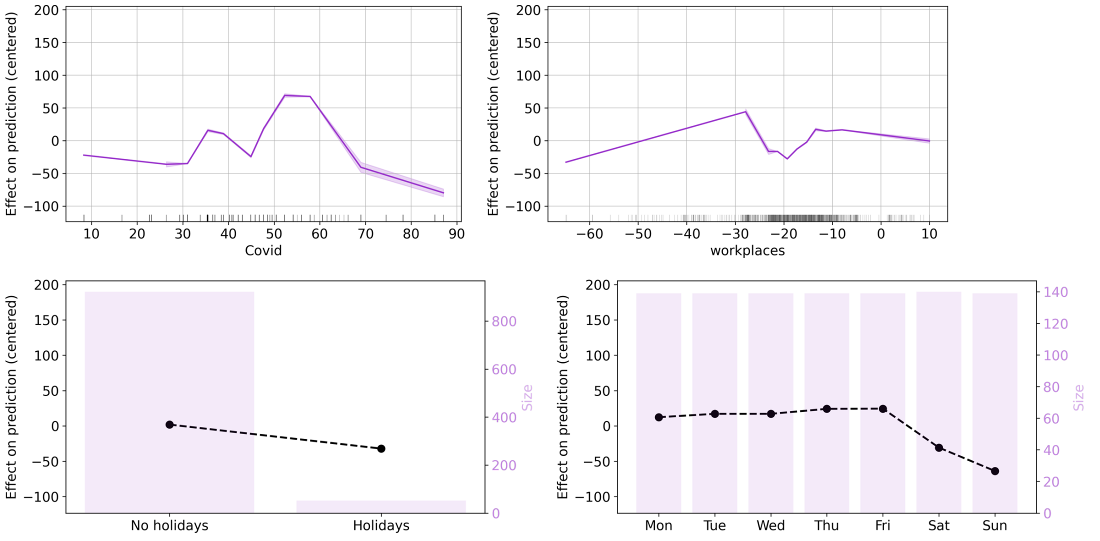


Figure S26. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

Russia: Validation curves

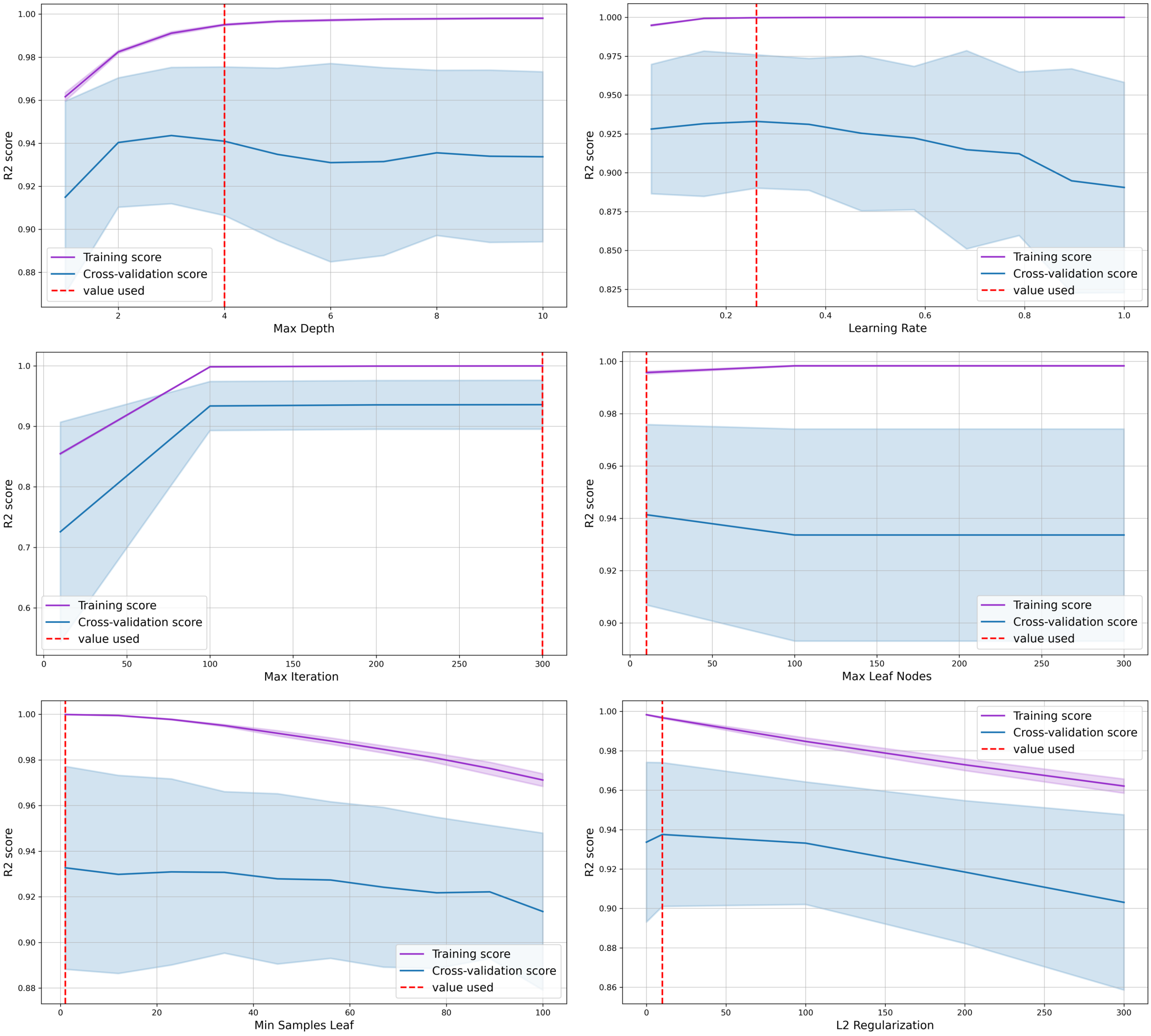


Figure S27. This figure displays validation curves for the hyperparameter of the Gradient Boosting, which is the best model in Russia. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

South Africa: Input data

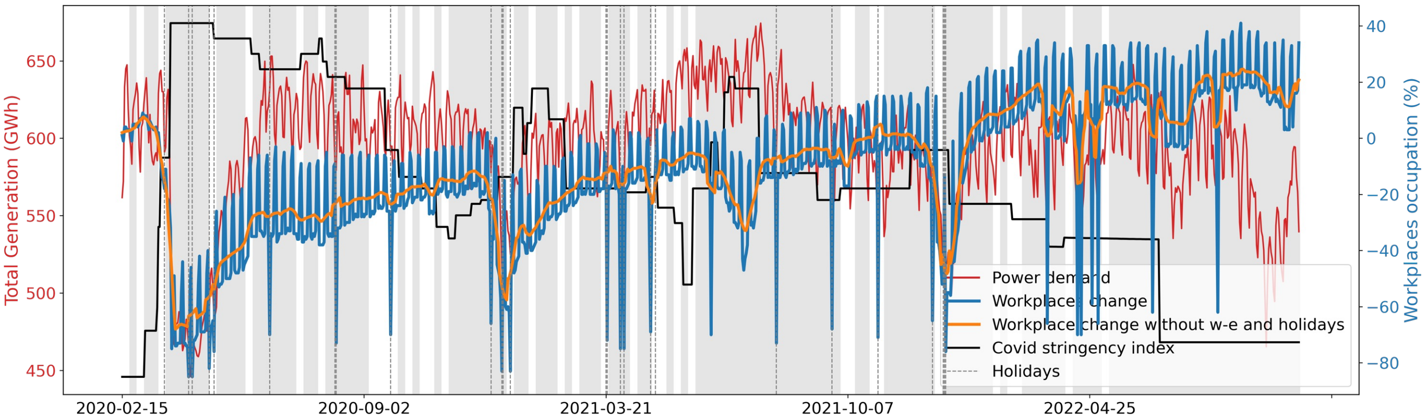


Figure S28. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

South Africa: models performance

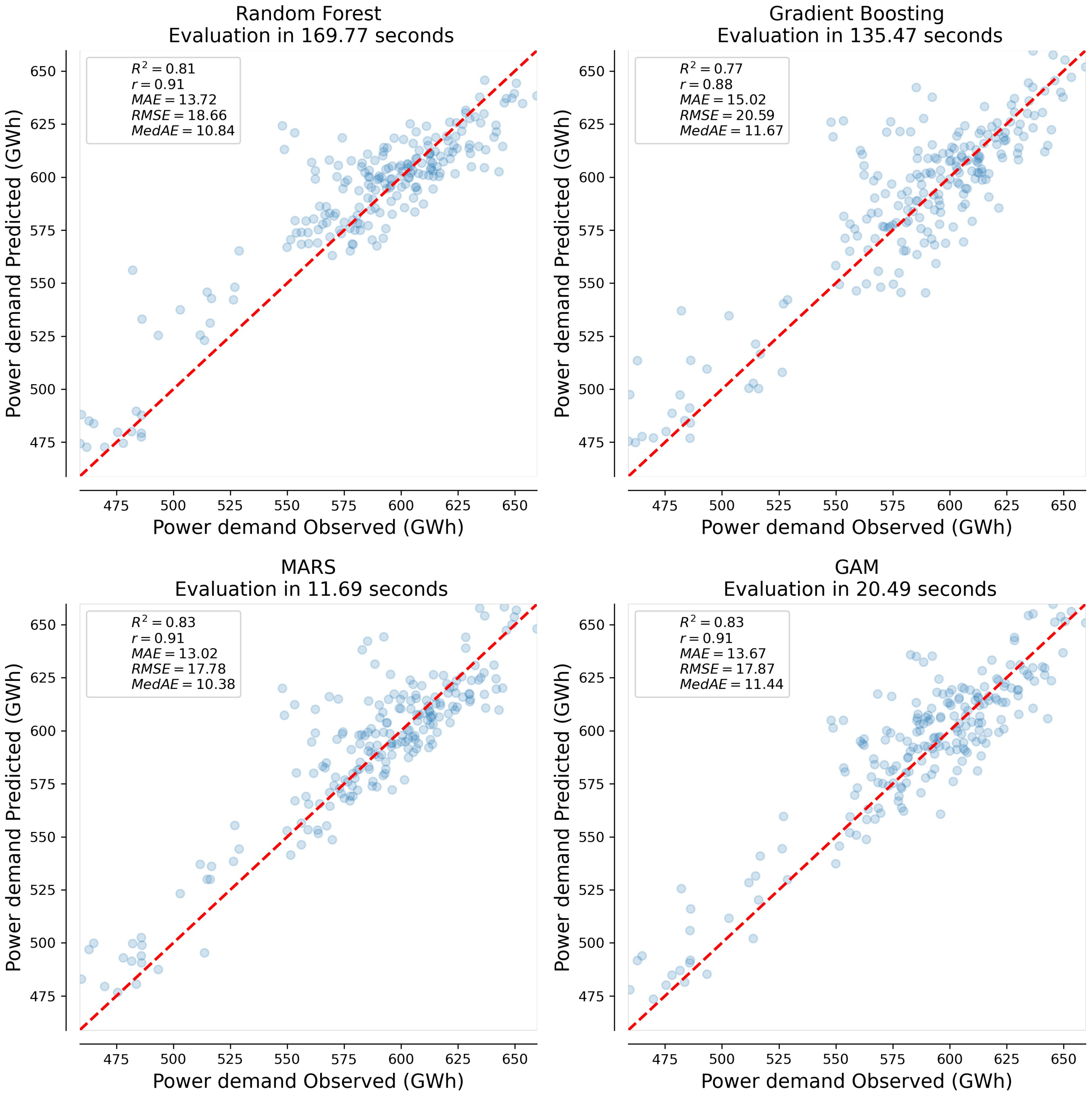


Figure S29. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

South Africa: feature permutation importance

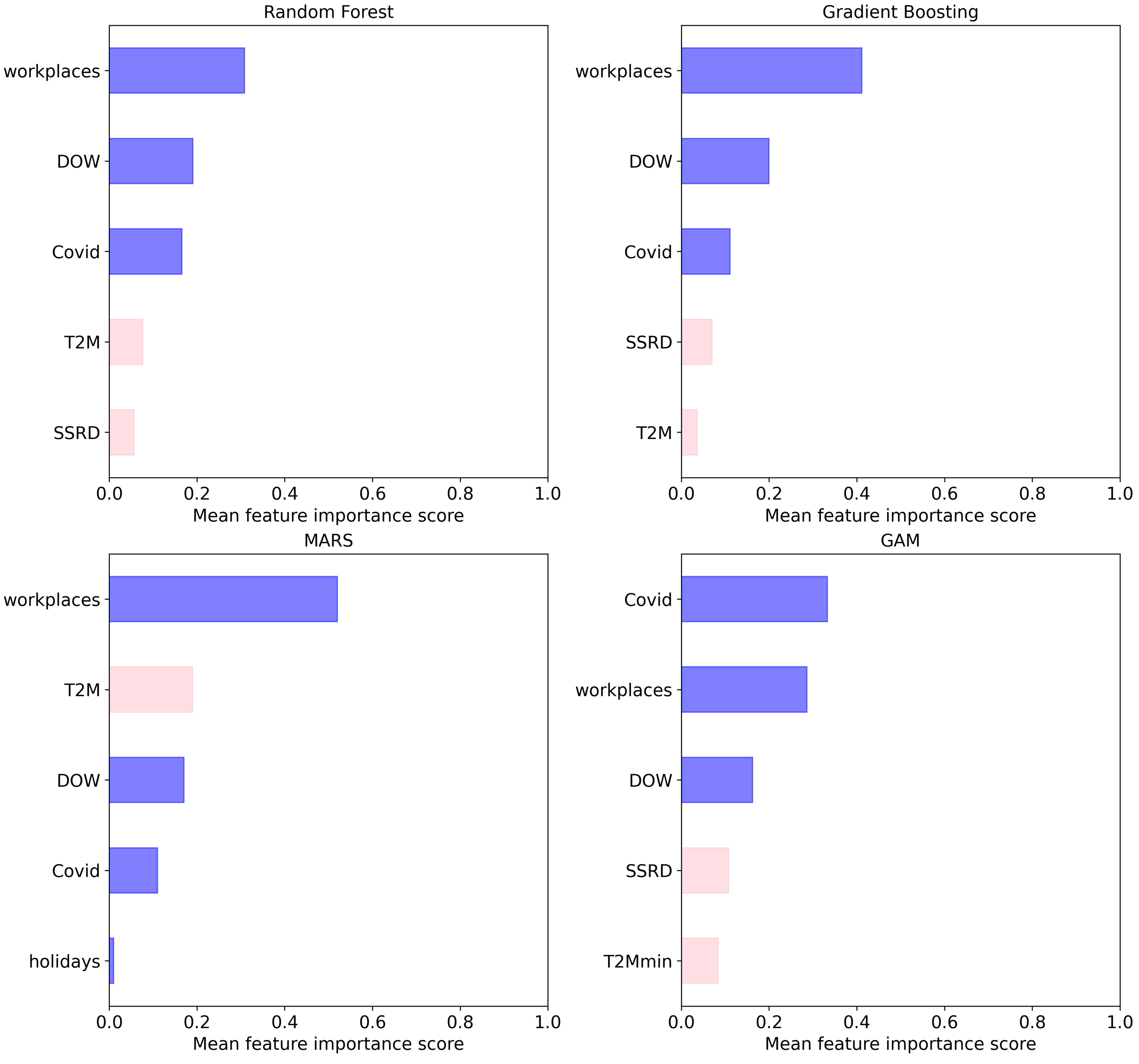
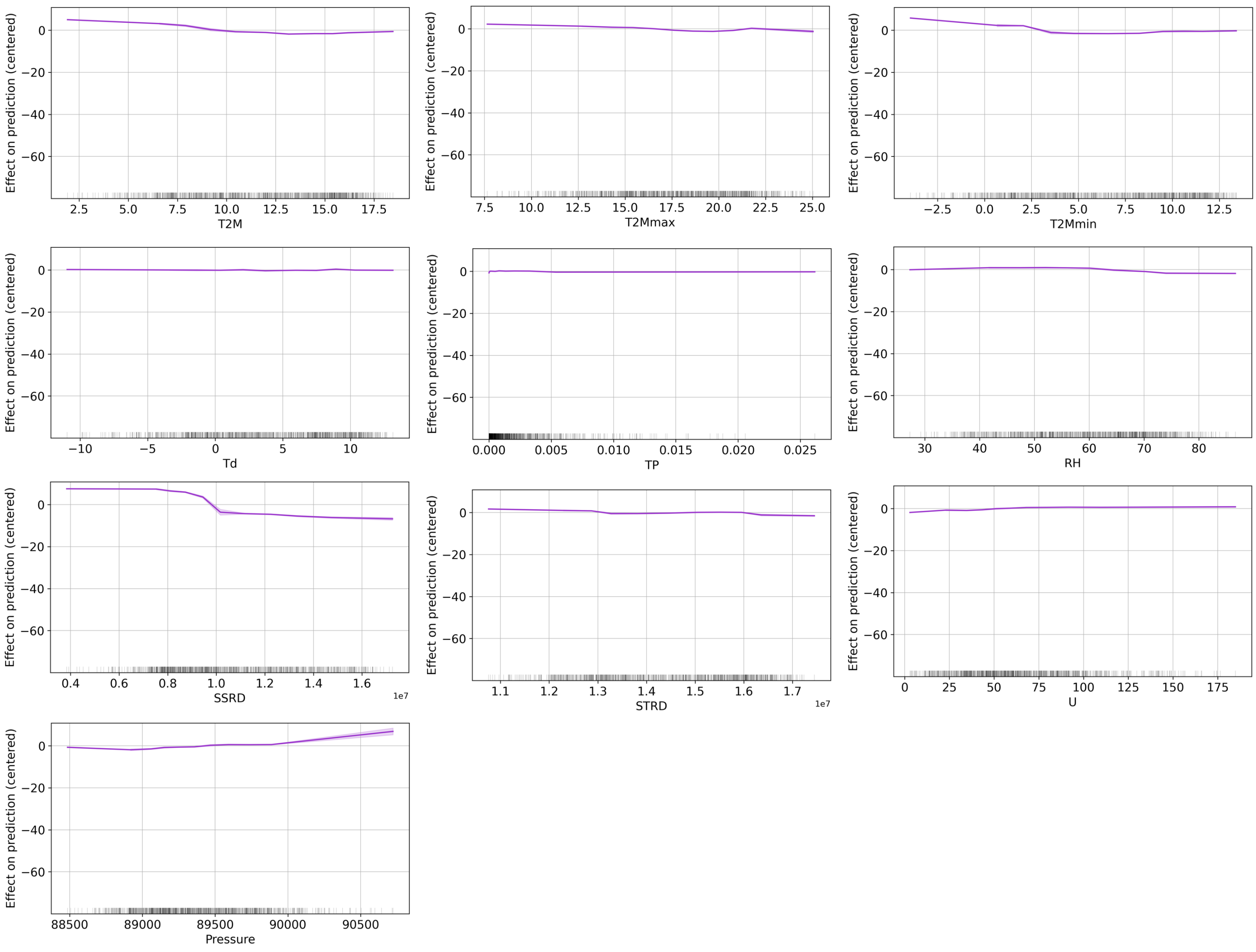


Figure S30. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

South Africa: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

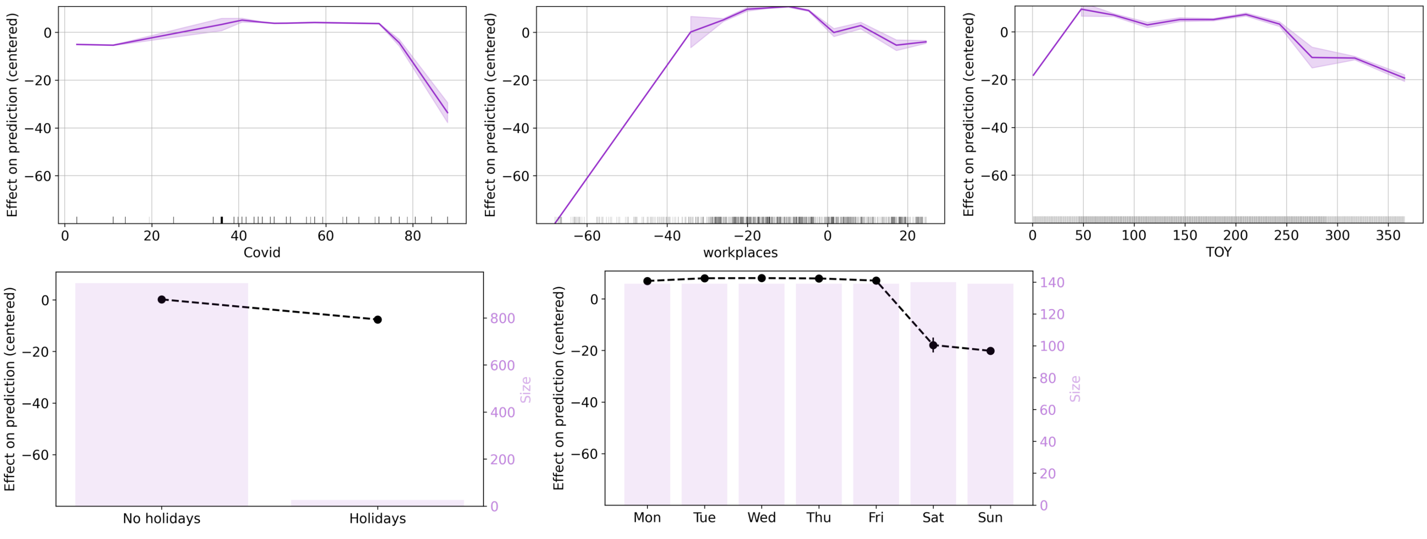


Figure S31. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.

South Africa: Validation curves

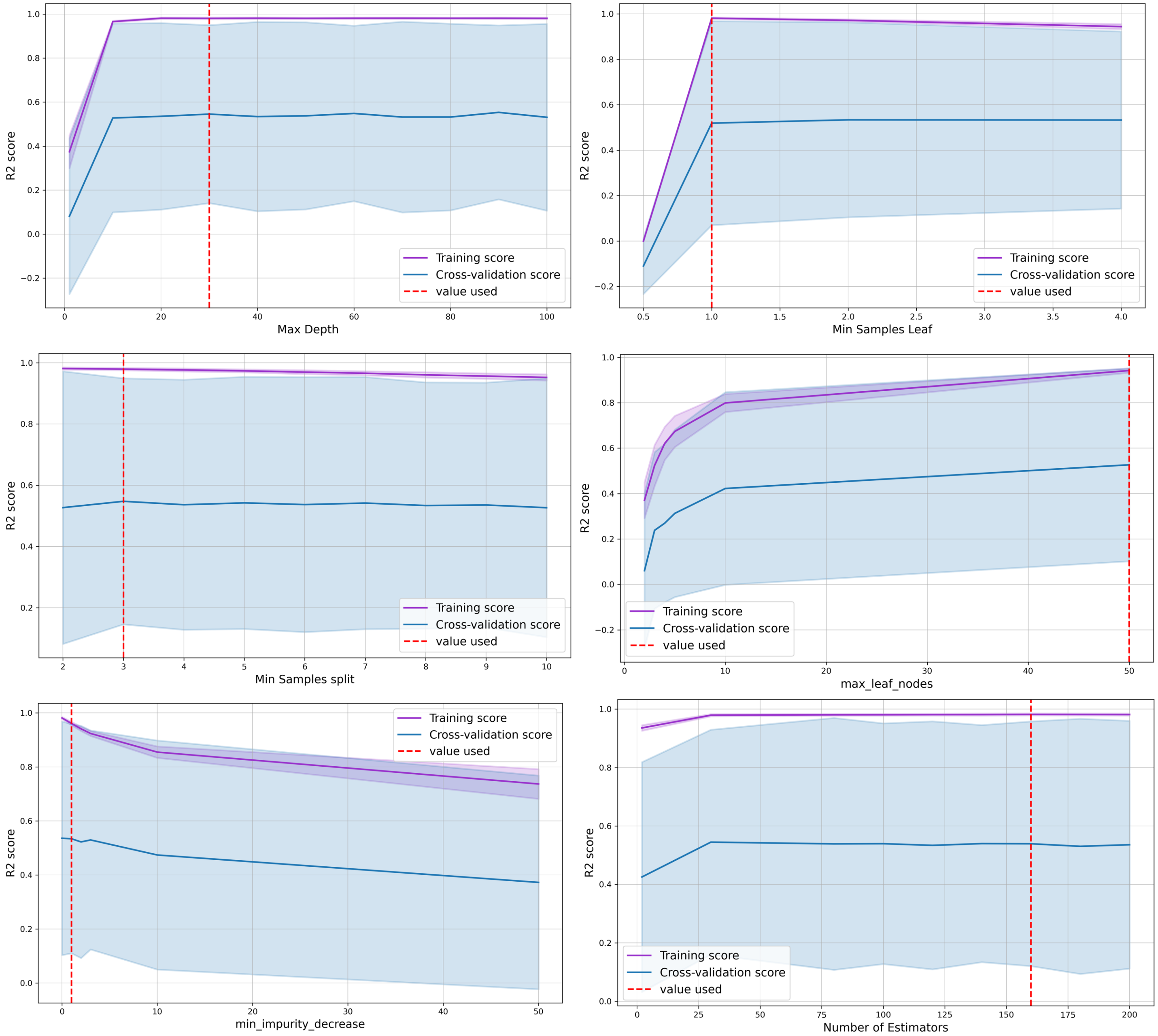


Figure S32. This figure displays validation curves for the hyperparameter of the GAM model, which is the best model in South Africa. The curves demonstrate how changes in the values of the hyperparameters affect the performance of the model, as measured by the R^2 score. The x-axis represents the range of values for the hyperparameter, and the y-axis shows the mean R^2 cross-validation score and R^2 training score. The validation and training scores are averaged over five scores calculated through cross-validation. The shaded areas represent the 95% confidence intervals for each curve. The red dashed line indicates the value of the hyperparameter selected during the grid-search process.

United States: Input data

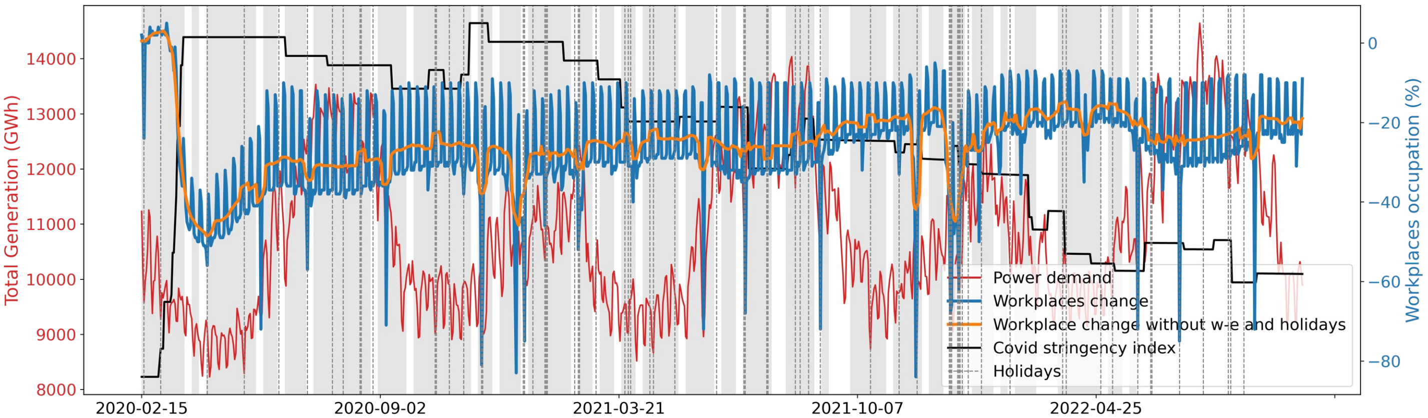


Figure S33. Evolution of human activity predictive features and power demand over the model training and testing period. Shaded area represents the train periods, blank area the test periods.

United States: models performance

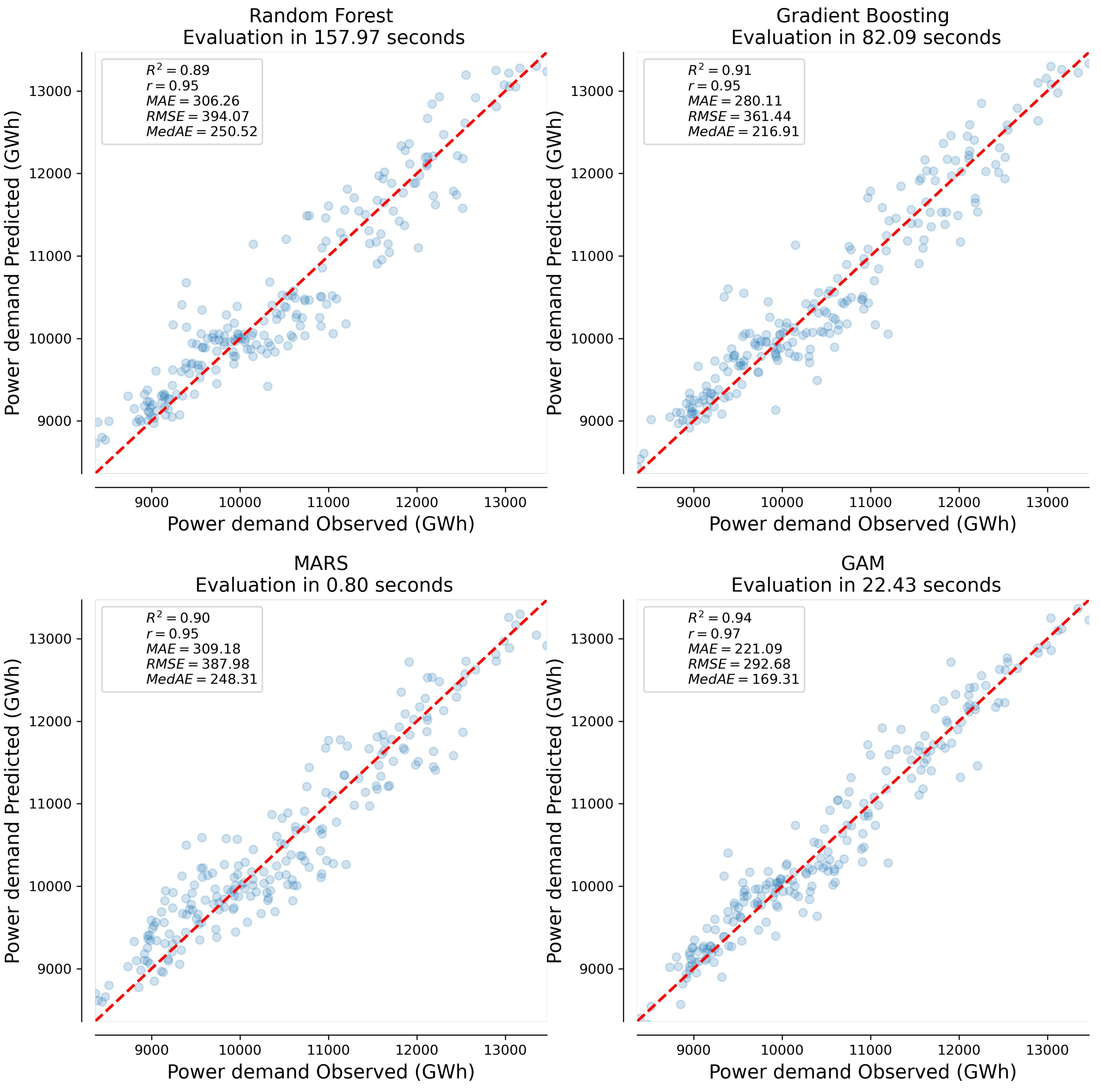


Figure S34. Comparison of machine learning model performance: predicted power demand plotted against observed power demand (blue points). The red dashed line represents the 1:1 line of perfect agreement between predictions and observations.

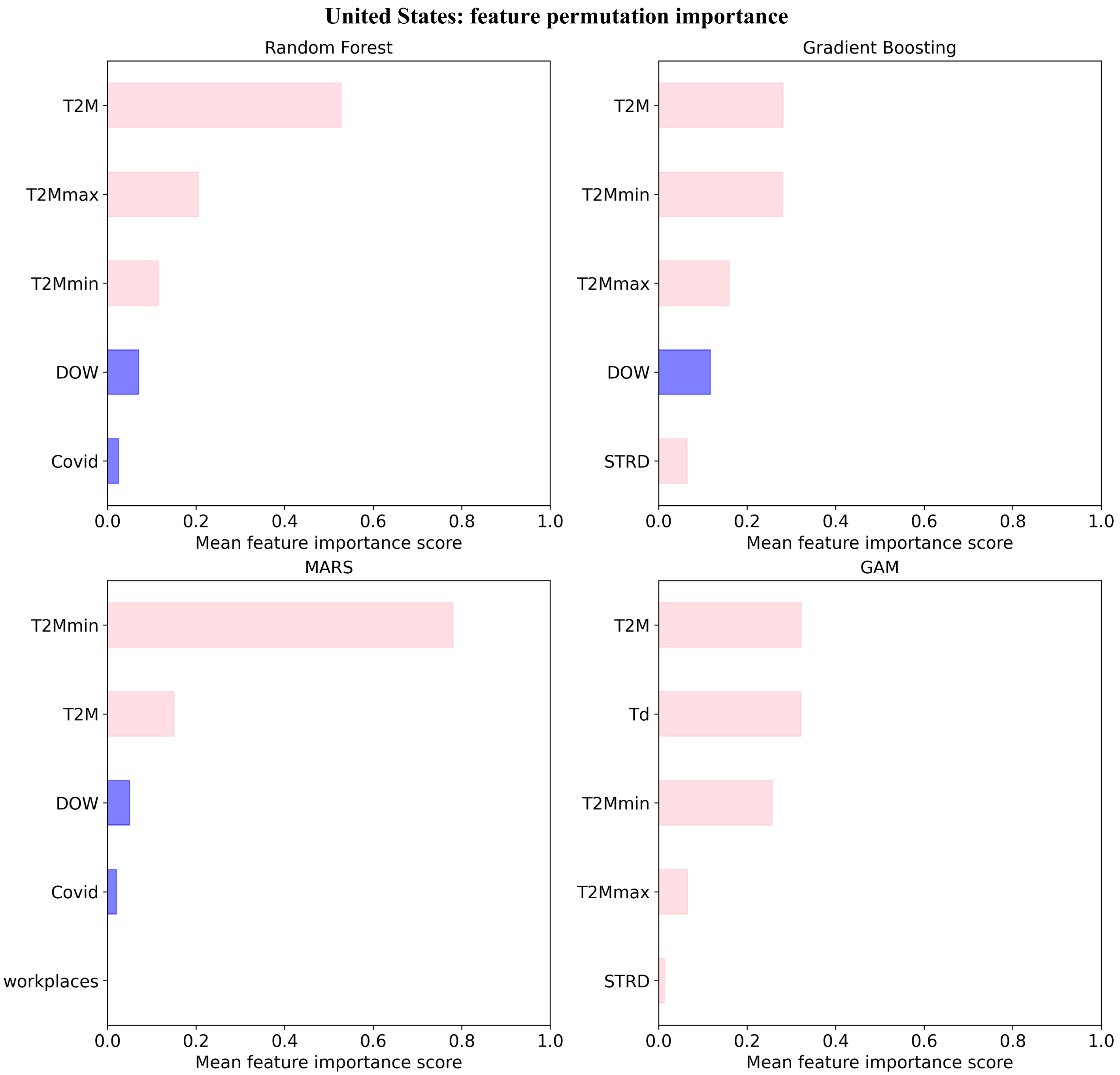
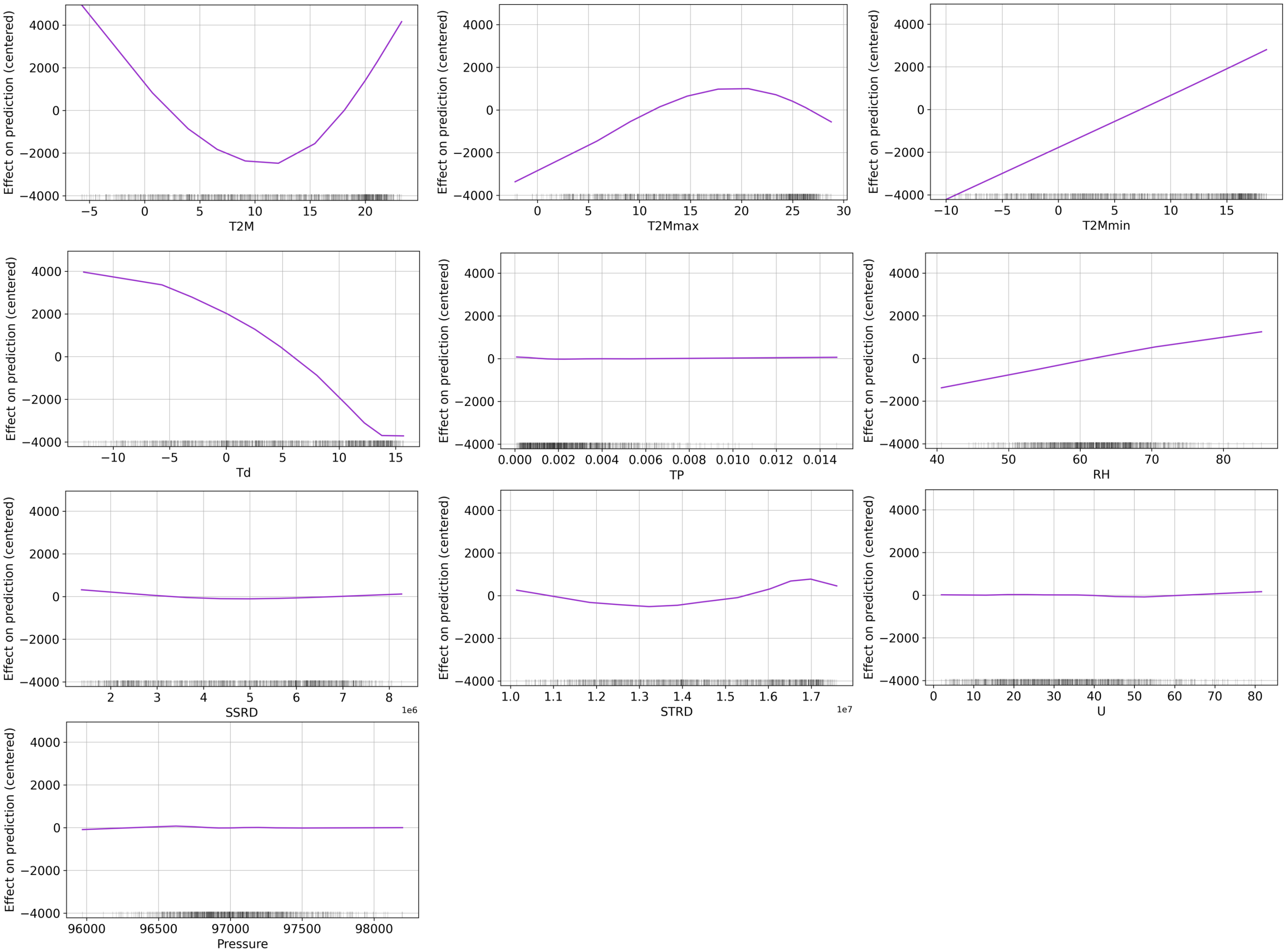


Figure S35. Permutation feature importance scores for the five most important predictive features for four different machine learning models: Random Forest, Gradient Boosting, Multivariate Adaptive Regression Splines (MARS), and Generalized Additive Models (GAM). The x-axis represents the countries, and the y-axis the different predictive features used in the models.

United States: ALE plots

a. Climate Predictive Features



b. Human Activity Predictive Features

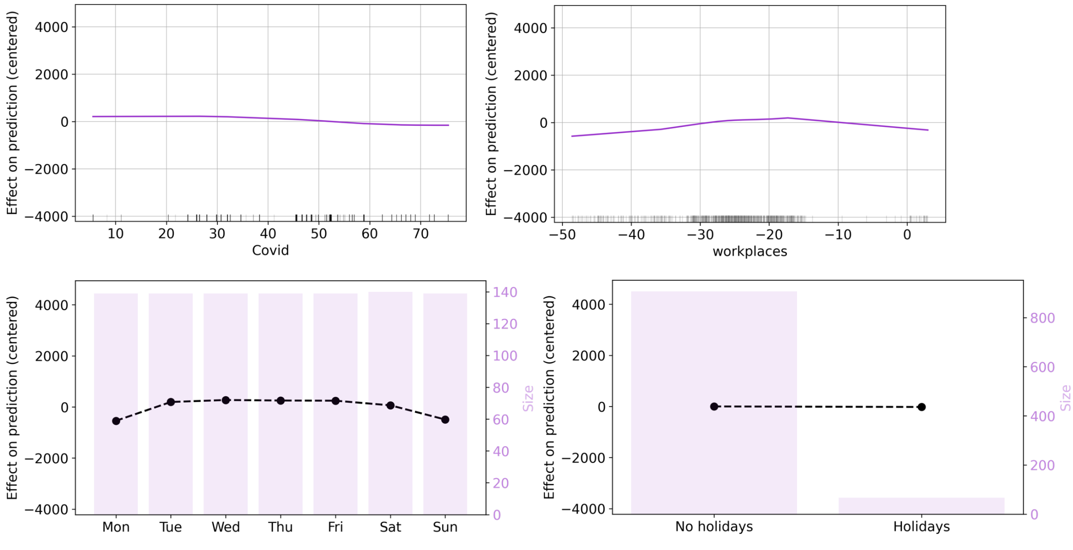


Figure S36. ALE plots depicting the effect of different predictive features on the target variable. The features are divided into two categories: (a) climate features and (b) human activity features. Each ALE plot shows the partial dependence of the target variable on a single feature while controlling for the effects of all other features. The x-axis represents the range of values for each feature, and the y-axis represents the corresponding change in the predicted value of the target variable. The shaded areas represent the 95% confidence intervals for each ALE curve.