Predictability Model of the Sea Ice Extent from Machine Learning

Young Eun Jeon[a] • Suk-Bok Kang[a] • Jung-In Seo[b]

[a] Department of Statistics, Yeungnam University, Gyeongsan, Korea
[b] Department of Information Statistics, Andong National University, Andong, Korea

1. Abstract

Background

- Tree-based machine learning techniques including random forest and extreme gradient boosting have been applied to various fields due to superior predictive performance.
- However, they have a disadvantage of not being able to capture trends in time series data.

Purpose

- To overcome this drawback, we propose a hybrid strategy based on the combination of tree-based machine learning and statistical techniques.
- In addition, Fourier terms are considered as a feature to handle a seasonal variation.

Analysis

- The superiority of the proposed strategy is demonstrated by a practical application using data for Arctic sea ice extent which is an important indicator showing the effect of global warming.

2. Models for Hybrid Strategy

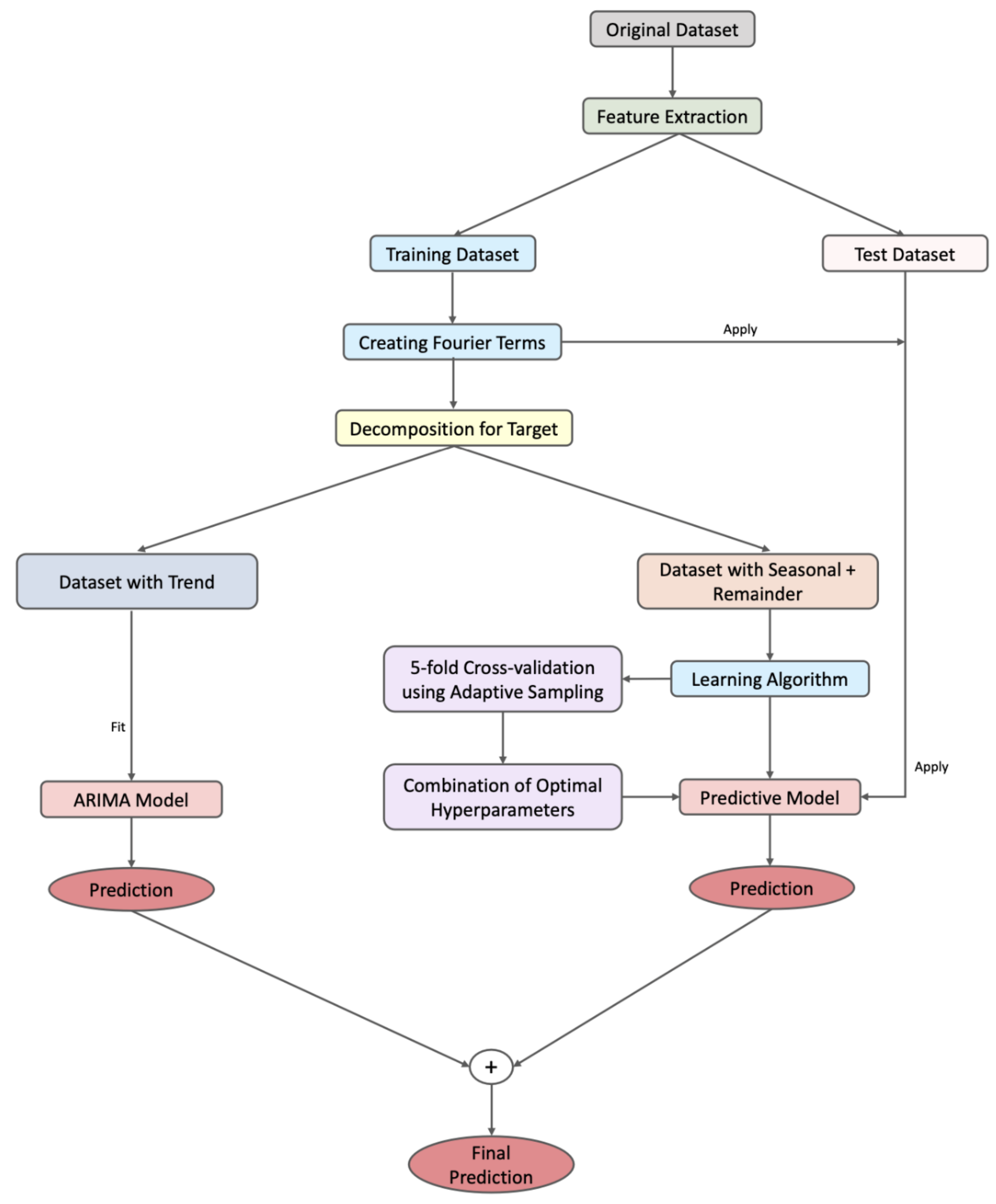


Figure 1. Prediction framework of the machine learning technique based on the proposed hybrid strategy

2.1 Trend Component

- **Autoregressive Integrated Moving Average (ARIMA) Model**

$$(1-\phi_1 B_1 - \cdots - \phi_p B_p)(1-B)^d Y_t = (1+\theta_1 B_1 + \cdots + \theta_q B_q)\epsilon_t, \epsilon_t \sim WN(0,\sigma_\epsilon^2)$$

✓ p : The order of the autoregression term
✓ d : The number of differences required to make a stationary time series
✓ q : The order of the moving average term

2.2 Detrend (Seasonal + Remainder) Component

- **Random Forest (RF)**

✓ RF creates multiple decision trees based on a bagging technique and combines their outputs to derive a single result.

- **Extreme Gradient Boosting (XGBoost)**

✓ XGBoost is an efficient implementation of the gradient boosting algorithm in terms of the performance and speed.

	Hyperparameter	Description
RF	mtry ntree nodesize	Number of features randomly selected as candidates at each split Number of trees to grow Minimum size of terminal nodes
XGBoost	nrounds eta max_depth gamma min_child_weight subsample colsample_bytree	Number of boosting iterations Learning rate Maximum depth of a tree Minimum loss reduction required to make a split Minimum sum of the instance weight needed in a child Subsample ratio of training data Ratio for subsampling of features

- **Stacking**

✓ Stacking is an ensemble machine learning (ML) algorithm that uses the prediction results generated by individual algorithms as the input data for the final model.
✓ This study considers XGBoost and a generalized linear model (GLM) as the final model.

3. Application

3.1 Data

- To demonstrate the superiority of the proposed strategy, this study uses sea ice extent data from an Arctic-wide change in sea ice between 1988-01-13 and 2021-08-16.
(Data Source : National Snow & Ice Data Center)
- The Arctic sea ice plays a role in controlling the global temperature by reflecting sunlight entering the Arctic.
- However, due to global warming, the sea ice extent in the Arctic continues to decrease, which can be seen in Figure 2.

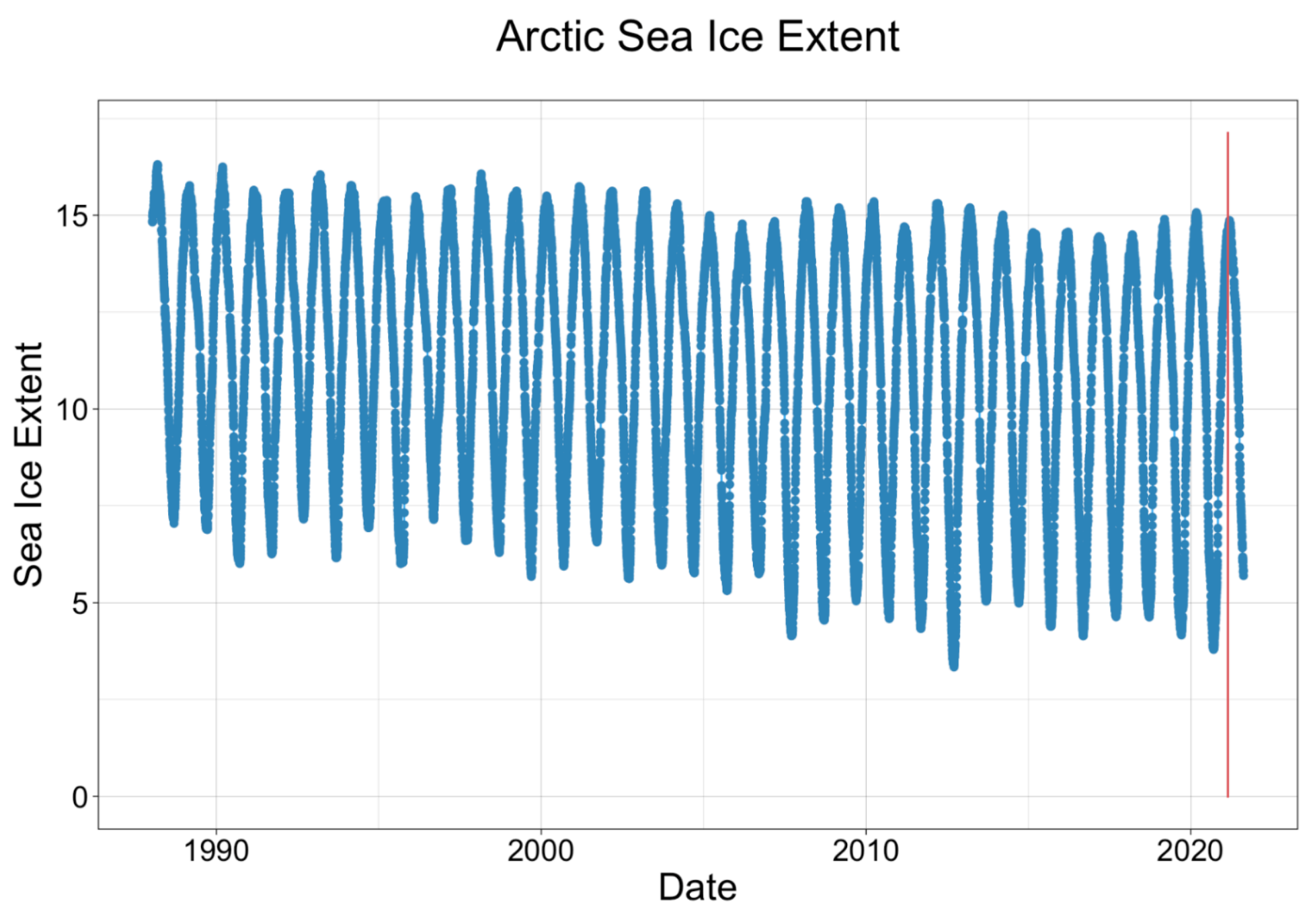


Figure 2. Arctic sea ice extent for the considered study period

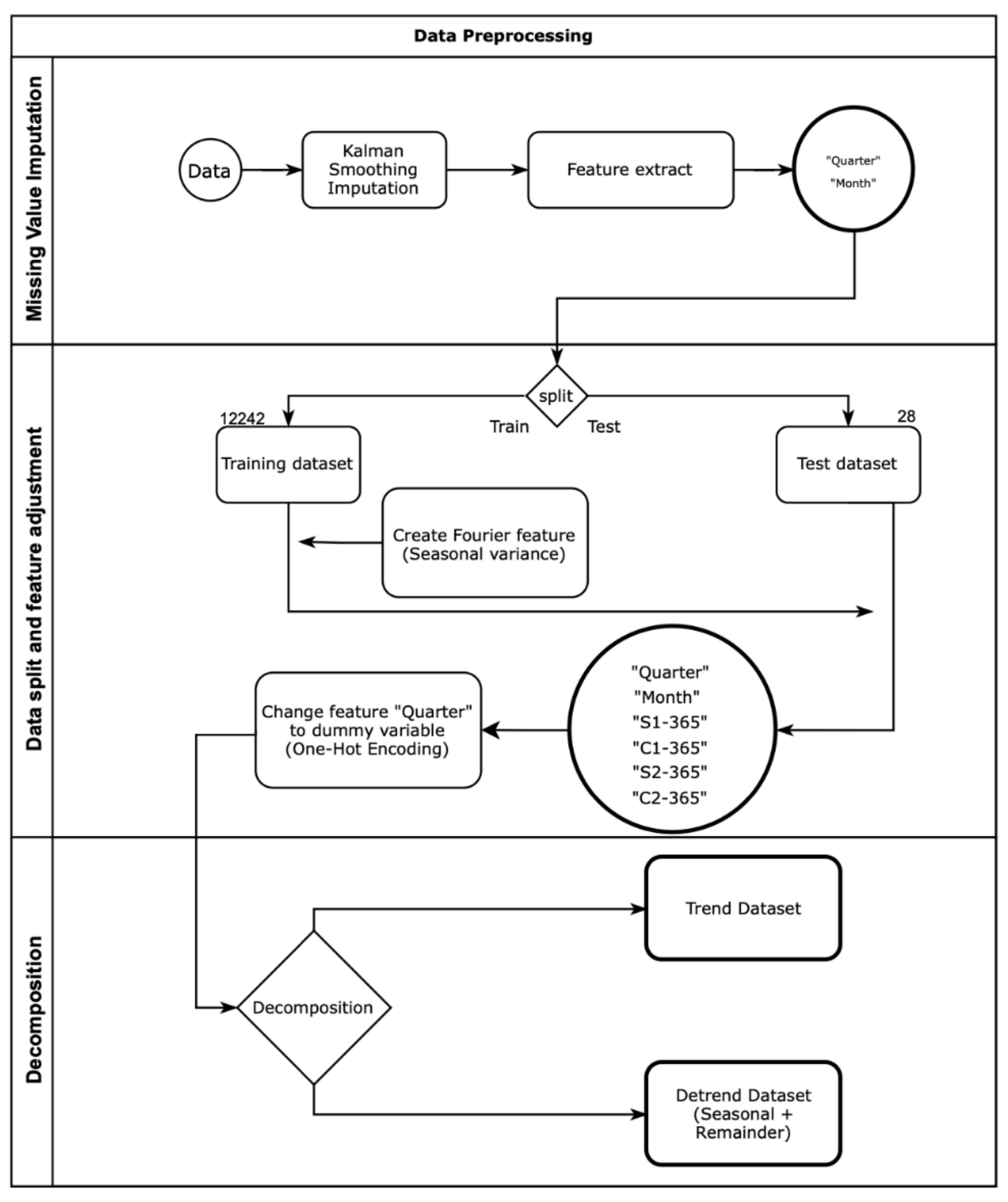

Data Preprocessing
Missing Value Imputation
Data
Kalman Smoothing Imputation
Feature extract
"Quarter" "Month"
Data split and feature adjustment
split
Train
Test
12242
28
Training dataset
Test dataset
Create Fourier feature (Seasonal variance)
"Quarter" "Month" "S1-365" "C1-365" "S2-365" "C2-365"
Change feature "Quarter" to dummy variable (One-Hot Encoding)
Decomposition
Decomposition
Trend Dataset
Detrend Dataset (Seasonal + Remainder)

Figure 3. Preprocessing flowchart

3.2 Results

- **Hyperparameter**

RF	XGBoost	Stacking	
		Stack.XGBoost	Stack.GLM
• mtry = 3 • ntree = 500 • nodesize = 5	• nrounds = 838 • eta = 0.0108 • max_depth = 8 • gamma = 4.0098 • min_child_weight = 5 • subsample = 0.3105 • colsample_bytree = 0.6982	• nrounds = 838 • eta = 0.0108 • max_depth = 8 • gamma = 4.0098 • min_child_weight = 5 • subsample = 0.3105 • colsample_bytree = 0.6982	• $\beta_0 = -0.0002$ • $\beta_{RF} = 0.4153$ • $\beta_{XGBoost} = 0.5912$

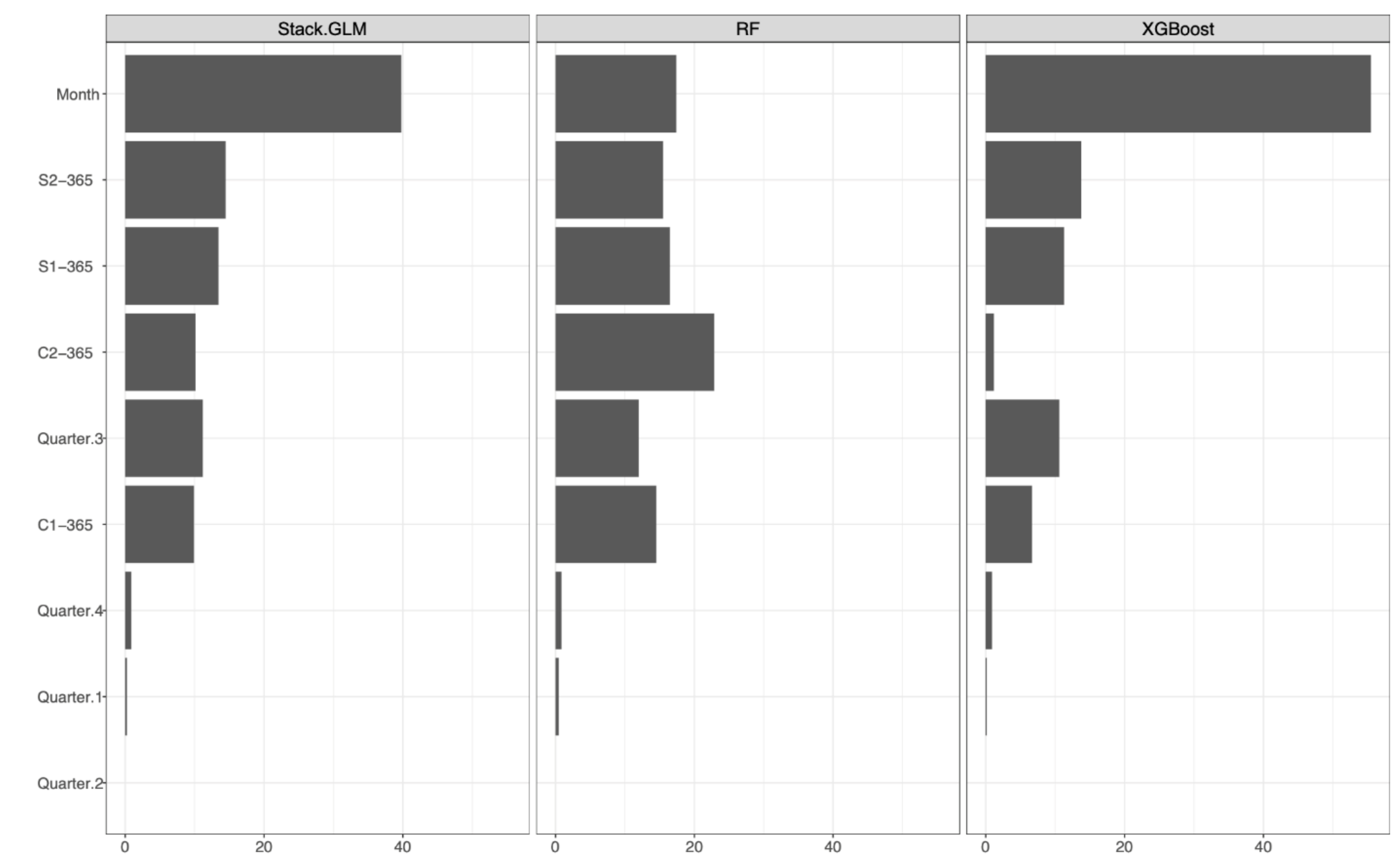


Figure 4. Feature importance based on model-specific

- **Model Evaluation**

	ME	RMSE	MAE	MPE	MAPE
RF	−0.180	0.217	0.186	−2.705	2.815
XGBoost	−0.186	0.220	0.189	−2.789	2.851
Stack.XGBoost	−0.143	0.200	0.170	−2.110	2.567
Stack.GLM	**−0.140**	**0.186**	**0.154**	**−2.072**	**2.305**

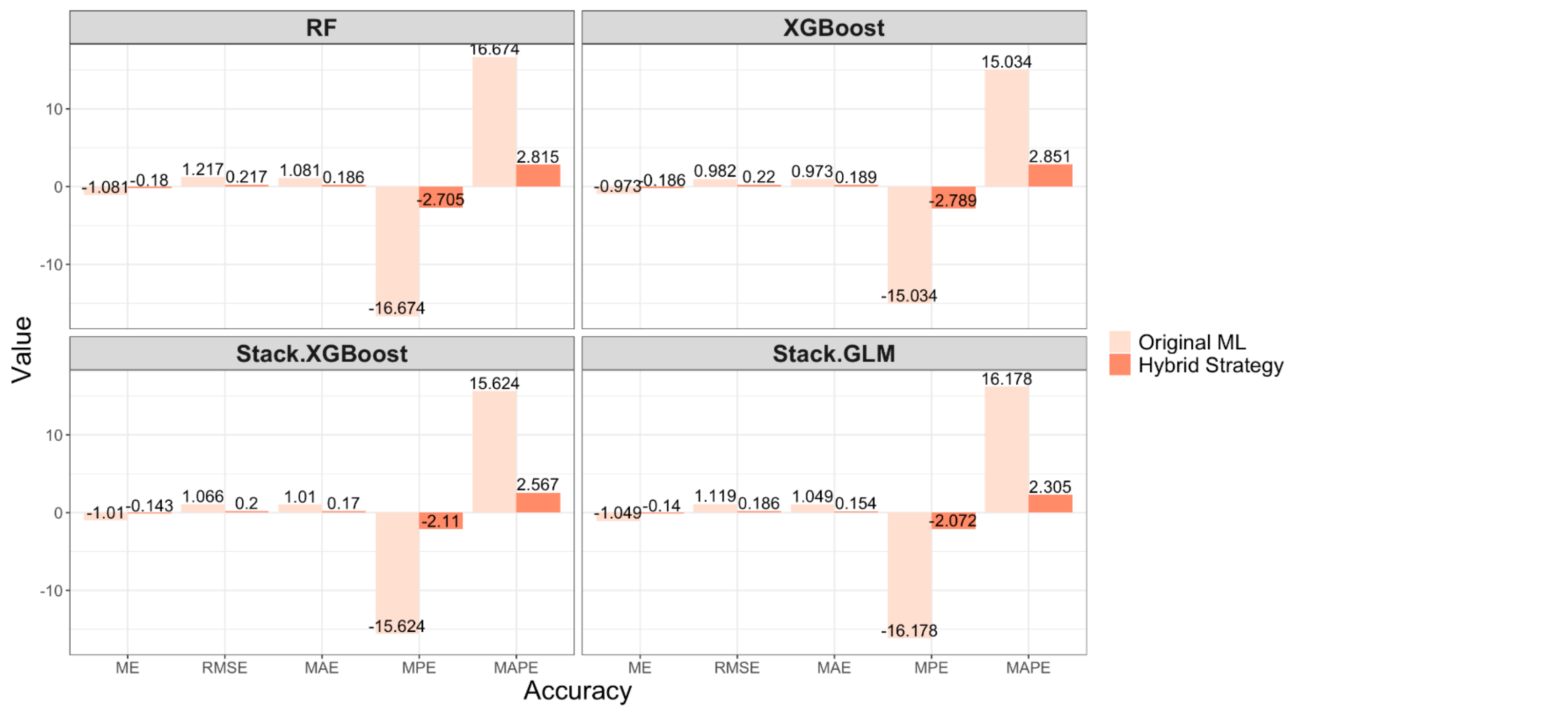
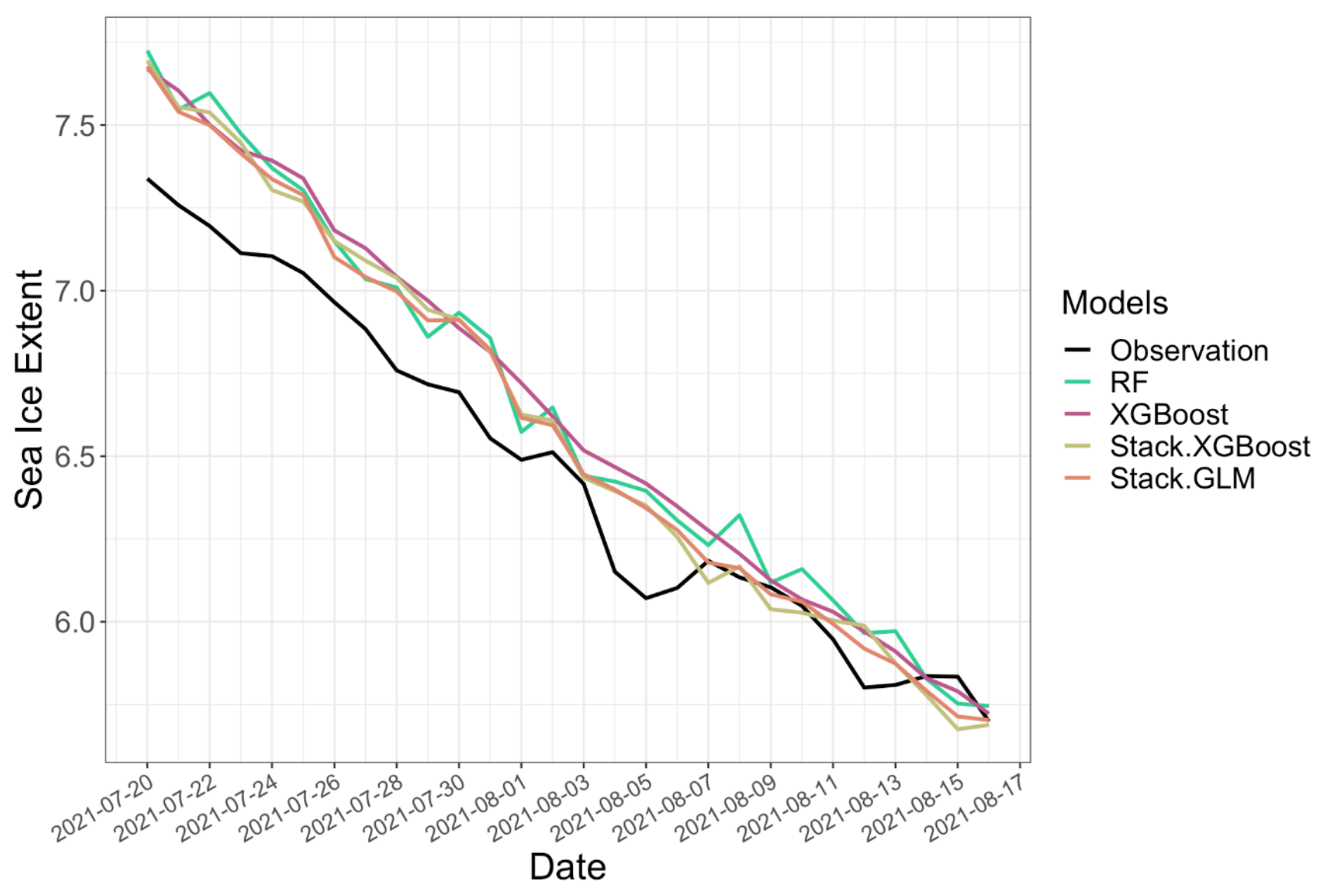
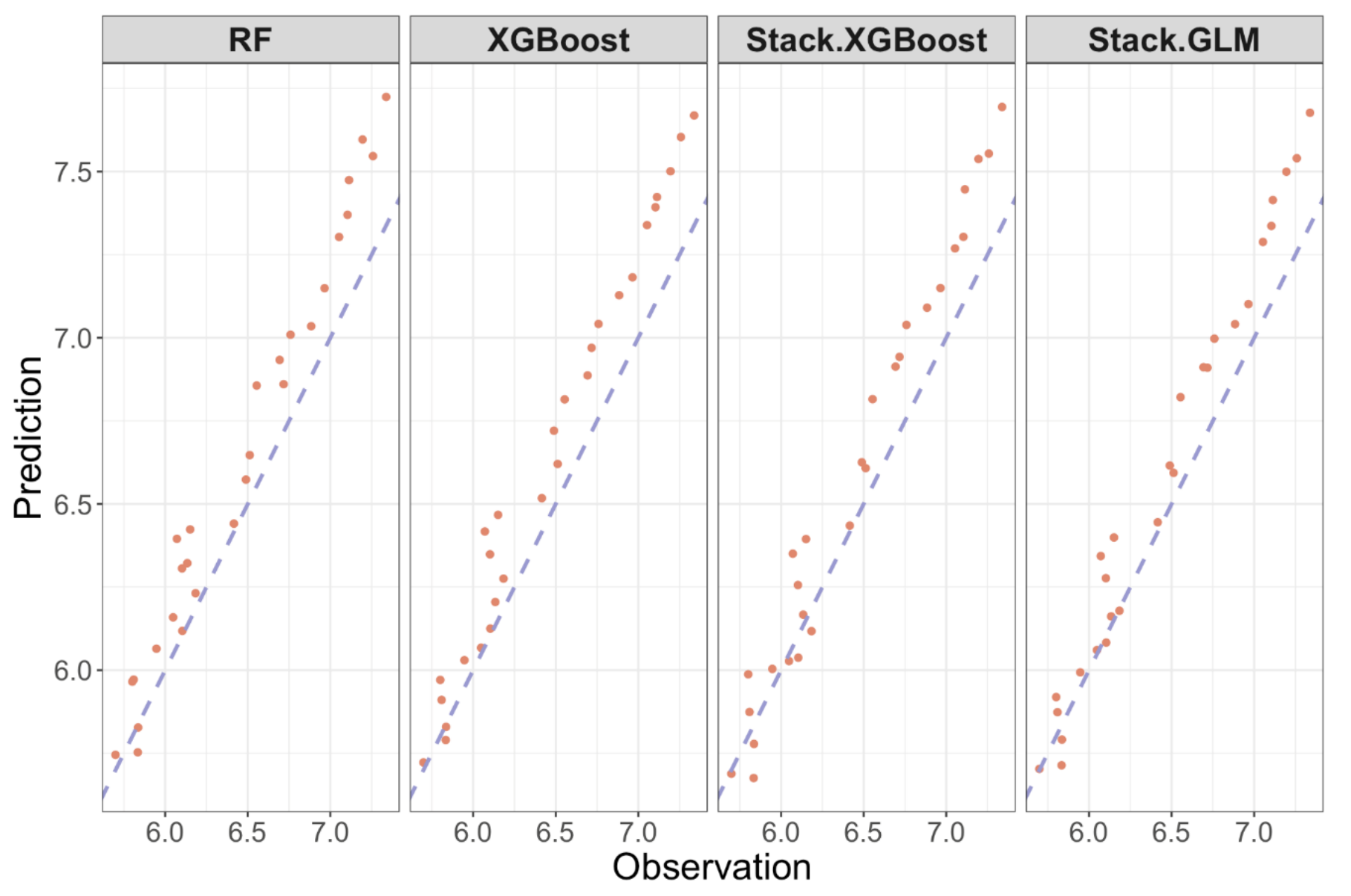


Figure 5. Comparison for accuracy between the original ML method and the proposed hybrid strategy

- **Plot between the Observation and Prediction**





4. Conclusion

- Our result shows that in a model-specific way, the ML models has high importance of the feature “Month”.
- In terms of the accuracy measure, Stack.GLM has the best predictive performance.
- The proposed hybrid strategy has better the prediction accuracy than the original ML method, which reveals that it will be a very good guideline for applying tree-based ML techniques to actual time series data.