CAAP Quarterly Report

September 25, 2024

Project Name: Determination of Potential Impact Radius for CO_2 Pipelines using Machine Learning Approach

Contract Number: 693JK32250011CAAP

Prime University: Texas A&M University

Prepared By: Sam Wang, qwang@tamu.edu, 979-845-9803

Reporting Period: 6/27/2024 – 9/26/2024

Project Activities for Reporting Period:

The following relevant tasks in the proposal have been completed:

- Studied the time for the scenario of concern to reach steady state. More details are provided in the appendix.
- Built machine learning models for all the geometries and concentrations to predict PIRs with CO_2 concentrations of 1%, 4%, and 9%. More details are provided in the appendix.

Project Financial Activities Incurred during the Reporting Period:

Based on the proposed budget, the cost is broken down into two parts:

- Efforts from the PI Dr. Wang for about 0.25 month.
- Efforts and work by graduate students, Chi-Yang Li and Jazmine Aiya D. Marquez, totally for about 3 months for each of them.

Project Activities with Cost Share Partners:

Dr. Wang's time and efforts (0.12 month) in this quarterly period are used as cost share. He devoted his time to supervise the graduate students, review all paperwork, discuss all simulation results, and prepare the progress report.

Project Activities with External Partners:

Dr. Wang officially participated in the ongoing **Skylark Joint Industry Project (JIP)**, led by DNV and UK HSE. On December 6, 2023, Dr. Wang officially submitted a statement of work, along with a budget and budget justification, to PHMSA for their consideration in joining the Skylark JIP. The extension of this work to continue with Skylark is still waiting for the official approval from PHMSA.

Potential Project Risks:

For the parametric study using Ansys Fluent, incorporating terrain information has increased the computation time. We have performed hundreds of CFD simulations which require a significant amount of time. With two PhD students working on this project, the simulations are successfully done along with the initial machine learning models.

Future Project Work:

- The future work is to study the evacuation time for appointed distance from the release point. Therefore, the emergency response plan can be organized accordingly to ensure the safety of the communities nearby.
- Conduct near-field simulations with the application of UDFs and UDRGMs in Ansys Fluent.
- Develop a web-based tool to determine the PIR for CO_2 pipelines and evacuation time for the surrounding public.

Potential Impacts to Pipeline Safety:

- The variables for pipeline characteristics and weather conditions cover the upper limits and lower limits of the current industrial practices; therefore, the machine-learning model is believed to have accurate predictions for other CO_2 pipelines in the range.

Appendix

1. Time to reach the steady state

The computational fluid dynamics simulations were performed at steady state. According to the discussions with the CO_2 pipeline operators, incidents involving CO_2 release from pipelines normally experience a discharge for about of 20 to 30 minutes. This study is to check the amount of time needed to reach the steady state.

To understand the time to reach steady state, a transient case with 0.1 s time step was conducted. The case with the farthest dispersion was used and the corresponding parameters are enumerated in Table 1. According to the simulation (Figure 1), the concentrations of 1%, 4%, and 9% reach steady state around 500, 180, and 80 seconds, which are much shorter than 20 minutes. Thus, simulations based on the steady state are rational.

Table 1. Parameters applied for study.

Variable	Pressure (MPa)	Diameter (inch)	Flow rate (MMcfd)	Wind speed (mph)	Temperature (°F)
Value	10	30	1300	25	60

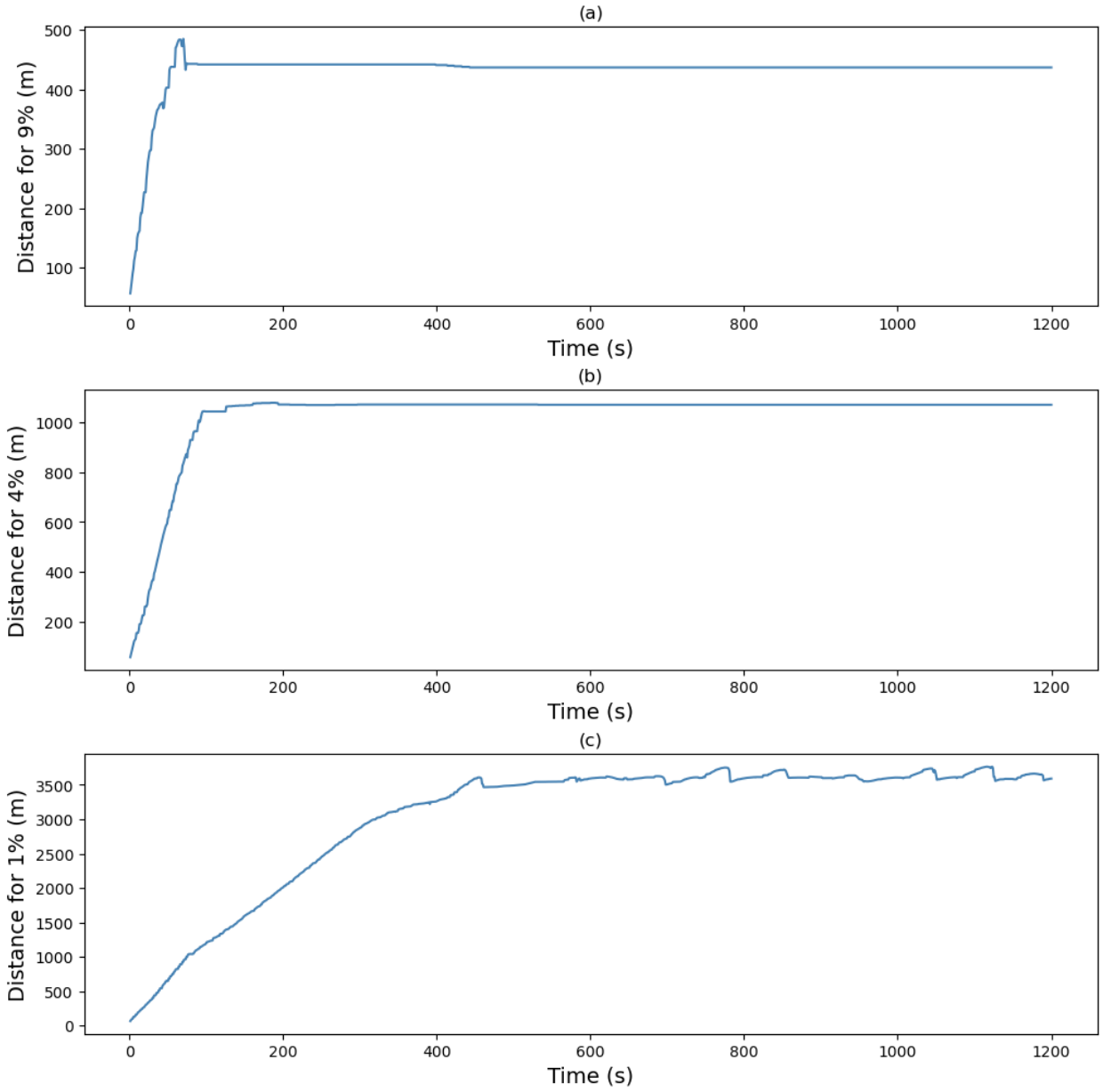

Figure 1. Distances of CO2 concentration versus time: (a) 9%, (b) 4%, and (c) 1%.

2. Machine learning models to predict PIRs

Because the distribution of the distances for the three different concentrations are quite divergent, we built three distinct models for each of them. The machine learning models applied for searching for the best model are multiple linear regression (MLR), Support Vector Regression (SVR), K nearest neighbors (KNN), random forest (RF), extreme gradient boosting regression (XGBoost), gradient boosting regression (GBR), and Bootstrap Aggregating (Bagging). R^2 with the 10-fold cross validation were used to search for the best model and to evaluate the performance of the models. In each model, the input (features) for the models are gauge pressure, diameter of pipeline, flow rate of CO_2, wind speed, and ambient temperature, and the output (response) is the corresponding distances from simulation.

With the random search of hyperparameters for each model, which is believed to be more efficient way to find the best model, the best version of each machine learning model is demonstrated in Tables 2-6. The predictions for 10-fold cross validations results are as Figure 2 to Figure 6. All the models hold higher than 0.93 on R^2, which represents high accuracy on prediction.

It is noted that the R-squared value represents the proportion of variance in the actual values that is explained by the model's predictions. An R-squared value of 1 indicates a perfect fit, where the predicted values exactly match the actual values. In this project, 10-fold cross-validation was used to optimize the model's hyperparameters and evaluate its performance. For each model, the R-squared was calculated by comparing the predicted and actual values across each fold, and the average R-squared from the 10 folds was computed to represent the model's overall performance. Even the worst-performing model among the 15 best models for each terrain of each concentration explained 93% of the variance, indicating a high degree of predictive accuracy, with many of them being very close to 1.

Table 2. Performance for each fine-tuned machine learning model for Flat.

CO_2 concentration (%)	Model	R^2	SD
9	Gradient Boosting	0.9665	0.0384
	Bagging	0.9691	0.0301
	Random Forest	0.9688	0.0300
	XGBoost	0.9782	0.0286
	K nearest neighbors	0.6703	0.1188
	Multiple Linear Regression	0.4806	0.1564
	Support Vector Regression	0.7775	0.0939
4	Gradient Boosting	0.9635	0.0270
	Bagging	0.9600	0.0355
	Random Forest	0.9604	0.0352
	XGBoost	0.9690	0.0453
	K nearest neighbors	0.7520	0.1533
	Multiple Linear Regression	0.5468	0.1470
	Support Vector Regression	0.7869	0.1187
1	Gradient Boosting	0.9849	0.0125
	Bagging	0.9833	0.0119
	Random Forest	0.9836	0.0105
	XGBoost	0.9886	0.0112
	K nearest neighbors	0.9242	0.0393
	Multiple Linear Regression	0.7943	0.0795
	Support Vector Regression	0.9236	0.0333

Table 3. Performance for each fine-tuned machine learning model for SH.

CO_2 concentration (%)	Model	R^2	SD
9	Gradient Boosting	0.9830	0.0348
	Bagging	0.9804	0.0124
	Random Forest	0.9806	0.0122
	XGBoost	0.9918	0.0093
	K nearest neighbors	0.6650	0.1575
	Multiple Linear Regression	0.4114	0.2990
	Support Vector Regression	0.7682	0.1172
4	Gradient Boosting	0.9672	0.0213
	Bagging	0.9663	0.0282
	Random Forest	0.9665	0.0251
	XGBoost	0.9700	0.0346
	K nearest neighbors	0.7345	0.1029
	Multiple Linear Regression	0.4470	0.1690
	Support Vector Regression	0.7738	0.0767
1	Gradient Boosting	0.9940	0.0039
	Bagging	0.9917	0.0043
	Random Forest	0.9918	0.0048
	XGBoost	0.9950	0.0026
	K nearest neighbors	0.9474	0.0263
	Multiple Linear Regression	0.7764	0.0841
	Support Vector Regression	0.9490	0.0292

Table 4. Performance for each fine-tuned machine learning model for BH.

CO_2 concentration (%)	Model	R^2	SD
9	Gradient Boosting	0.9875	0.0079
	Bagging	0.9794	0.0109
	Random Forest	0.9795	0.0110
	XGBoost	0.9878	0.0075
	K nearest neighbors	0.6632	0.0999
	Multiple Linear Regression	0.5376	0.1397
	Support Vector Regression	0.7669	0.0921
4	Gradient Boosting	0.9301	0.0409
	Bagging	0.9272	0.0698
	Random Forest	0.9301	0.0629
	XGBoost	0.9288	0.0545
	K nearest neighbors	0.6416	0.1877
	Multiple Linear Regression	0.2711	0.3867
	Support Vector Regression	0.7174	0.1020
1	Gradient Boosting	0.9605	0.0237
	Bagging	0.9566	0.0293
	Random Forest	0.9575	0.0271
	XGBoost	0.9627	0.0210
	K nearest neighbors	0.7759	0.1128
	Multiple Linear Regression	0.5947	0.1612
	Support Vector Regression	0.8121	0.0556

Table 5. Performance for each fine-tuned machine learning model for VM.

CO_2 concentration (%)	Model	R^2	SD
9	Gradient Boosting	0.9618	0.0364
	Bagging	0.9567	0.0298
	Random Forest	0.9574	0.0242
	XGBoost	0.9725	0.0220
	K nearest neighbors	0.6552	0.1310
	Multiple Linear Regression	0.4830	0.1194
	Support Vector Regression	0.7775	0.1117
4	Gradient Boosting	0.9160	0.0592
	Bagging	0.9232	0.0656
	Random Forest	0.9244	0.0624
	XGBoost	0.9330	0.0896
	K nearest neighbors	0.6489	0.1178
	Multiple Linear Regression	0.3963	0.0786
	Support Vector Regression	0.6946	0.1359
1	Gradient Boosting	0.9907	0.0092
	Bagging	0.9801	0.0091
	Random Forest	0.9801	0.0086
	XGBoost	0.9930	0.0054
	K nearest neighbors	0.8853	0.0364
	Multiple Linear Regression	0.7816	0.0701
	Support Vector Regression	0.8801	0.0260

Table 6. Performance for each fine-tuned machine learning model for VB.

CO_2 concentration (%)	Model	R^2	SD
9	Gradient Boosting	0.9656	0.0176
	Bagging	0.9713	0.0225
	Random Forest	0.9714	0.0228
	XGBoost	0.9762	0.0238
	K nearest neighbors	0.7500	0.1345
	Multiple Linear Regression	0.5813	0.1287
	Support Vector Regression	0.8231	0.1026
4	Gradient Boosting	0.9428	0.0444
	Bagging	0.9462	0.0345
	Random Forest	0.9480	0.0326
	XGBoost	0.9626	0.0264
	K nearest neighbors	0.7567	0.1161
	Multiple Linear Regression	0.4461	0.1998
	Support Vector Regression	0.7800	0.0954
1	Gradient Boosting	0.9942	0.0044
	Bagging	0.9859	0.0047
	Random Forest	0.9861	0.0047
	XGBoost	0.9952	0.0028
	K nearest neighbors	0.8901	0.0305
	Multiple Linear Regression	0.7897	0.0968
	Support Vector Regression	0.8940	0.0269

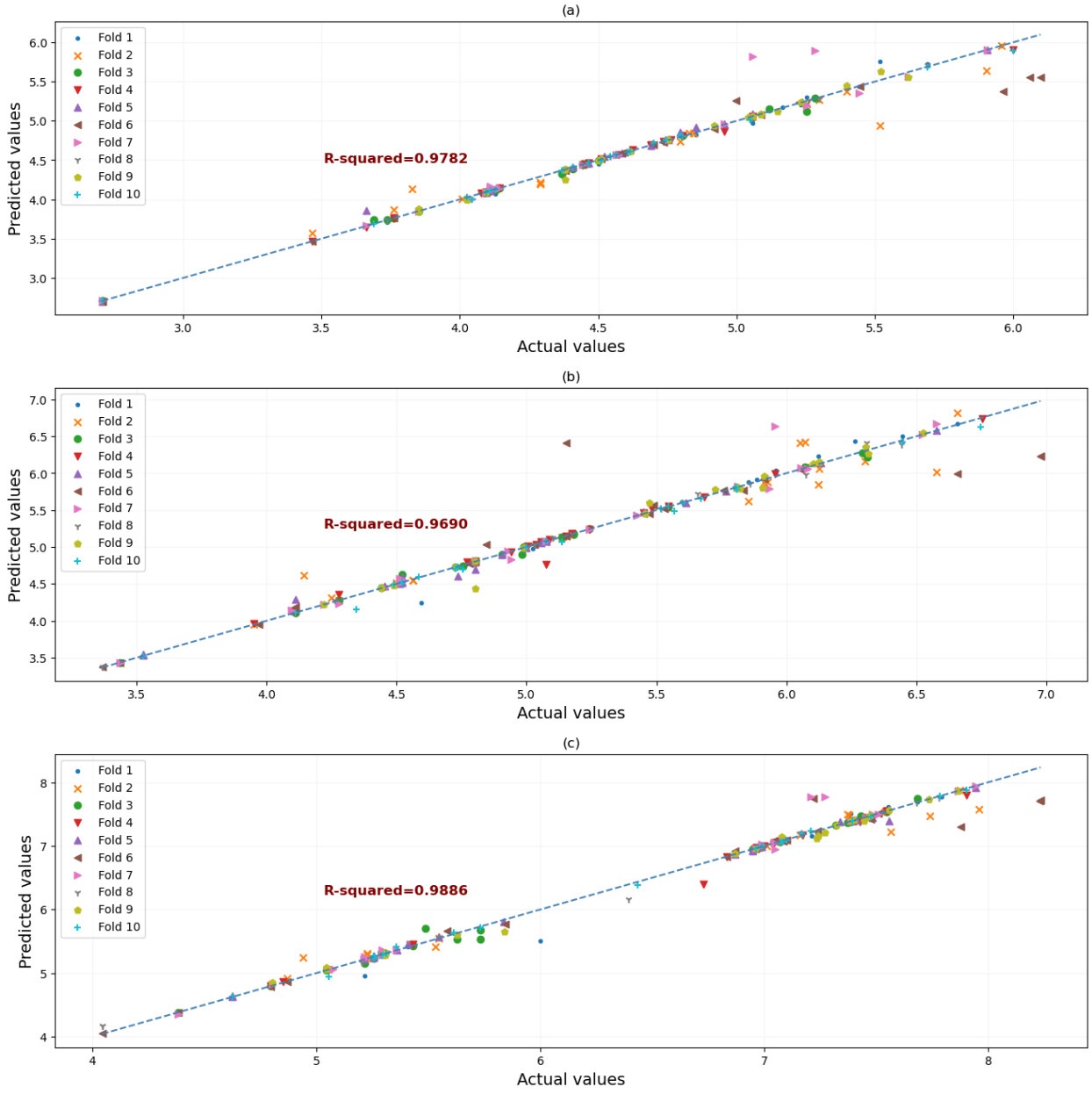

Figure 2. Actual vs. Predicted Values (10-fold cross validation) for Flat: (a) Distance for 9% CO_2, (b) Distance for 4% CO_2, and (c) Distance for 1% CO_2.

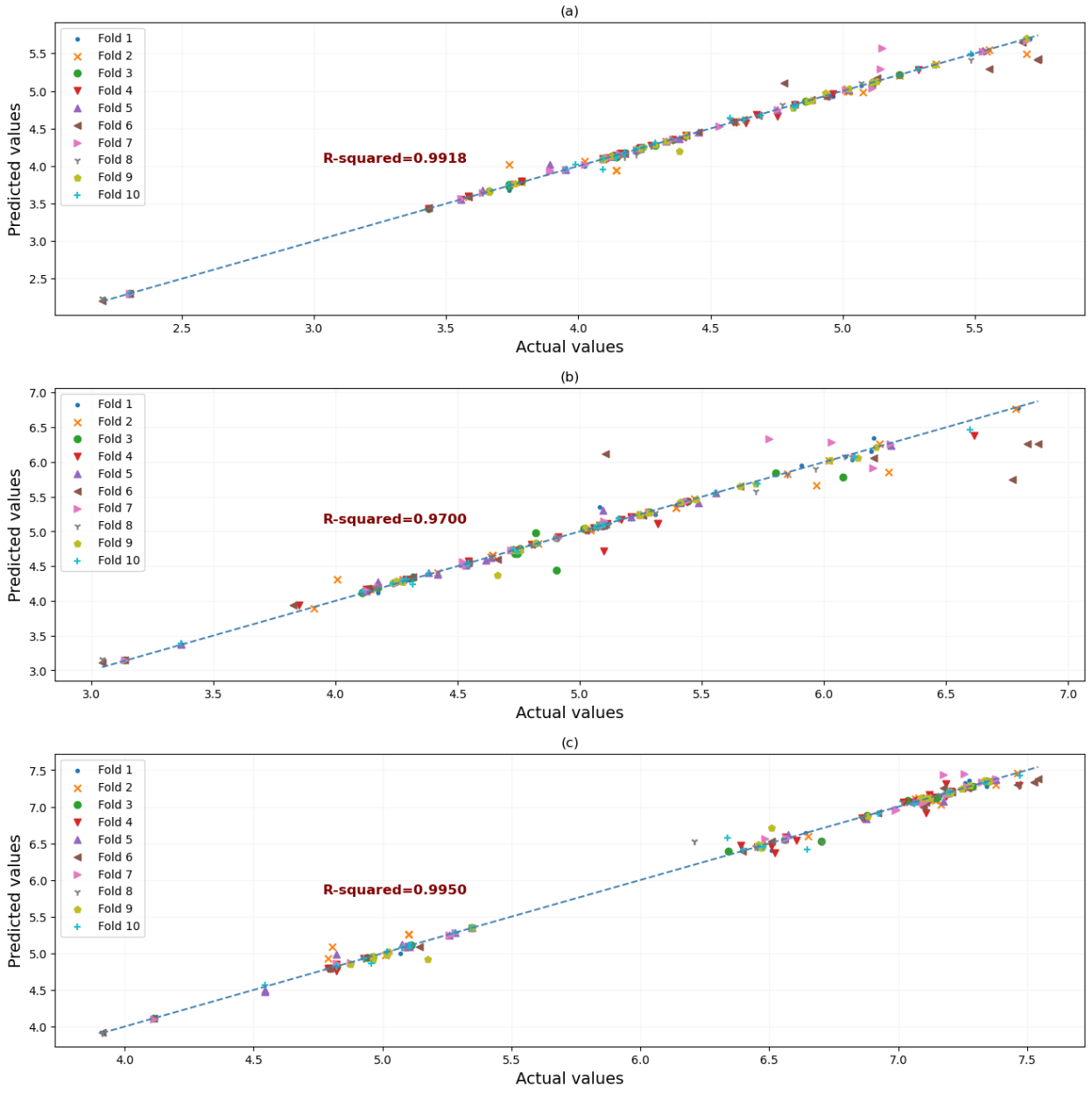

Figure 3. Actual vs. Predicted Values (10-fold cross validation) for SH: (a) Distance for 9% CO_2, (b) Distance for 4% CO_2, and (c) Distance for 1% CO_2.

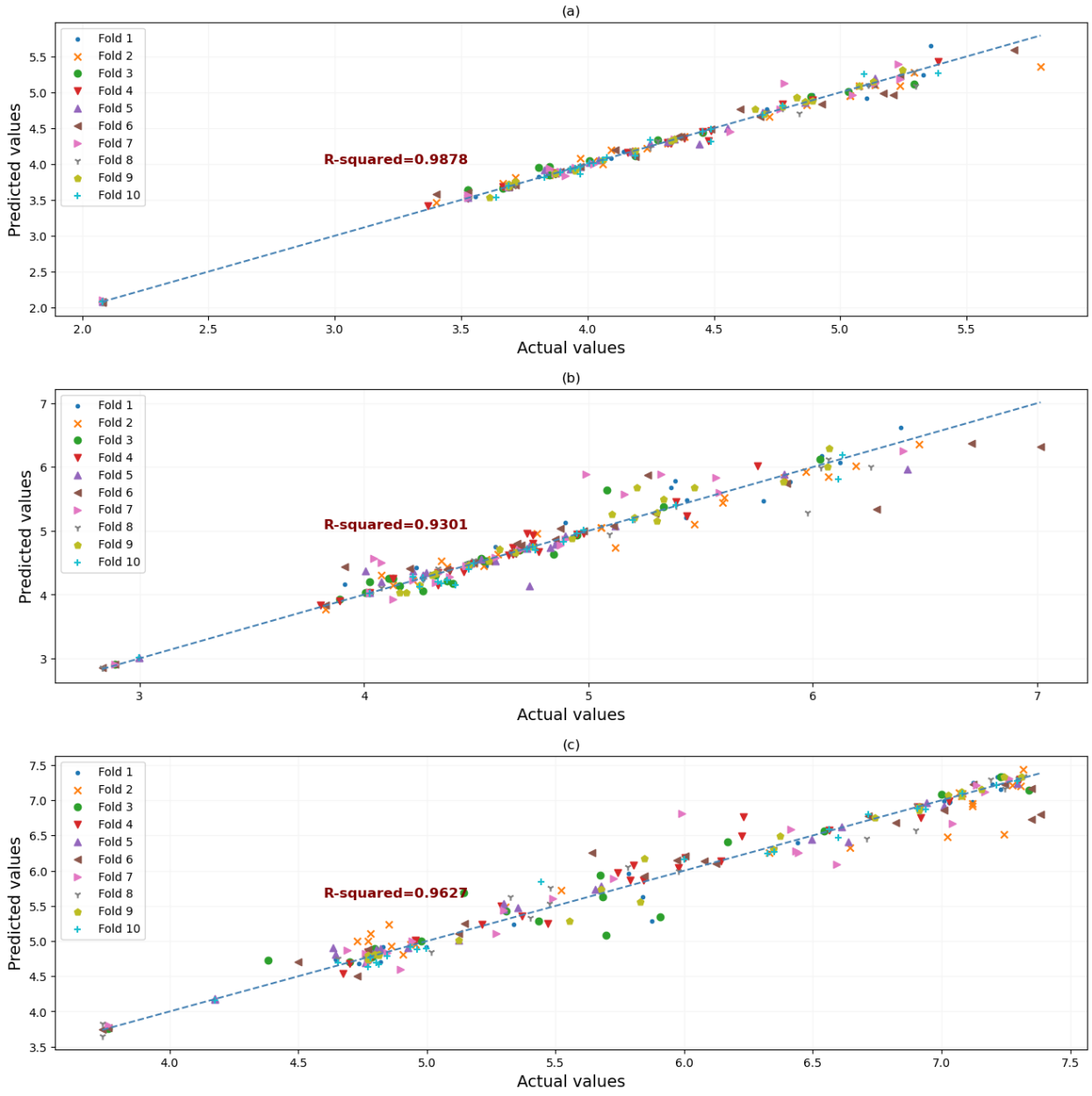

Figure 4. Actual vs. Predicted Values (10-fold cross validation) for BH: (a) Distance for 9% CO_2, (b) Distance for 4% CO_2, and (c) Distance for 1% CO_2.

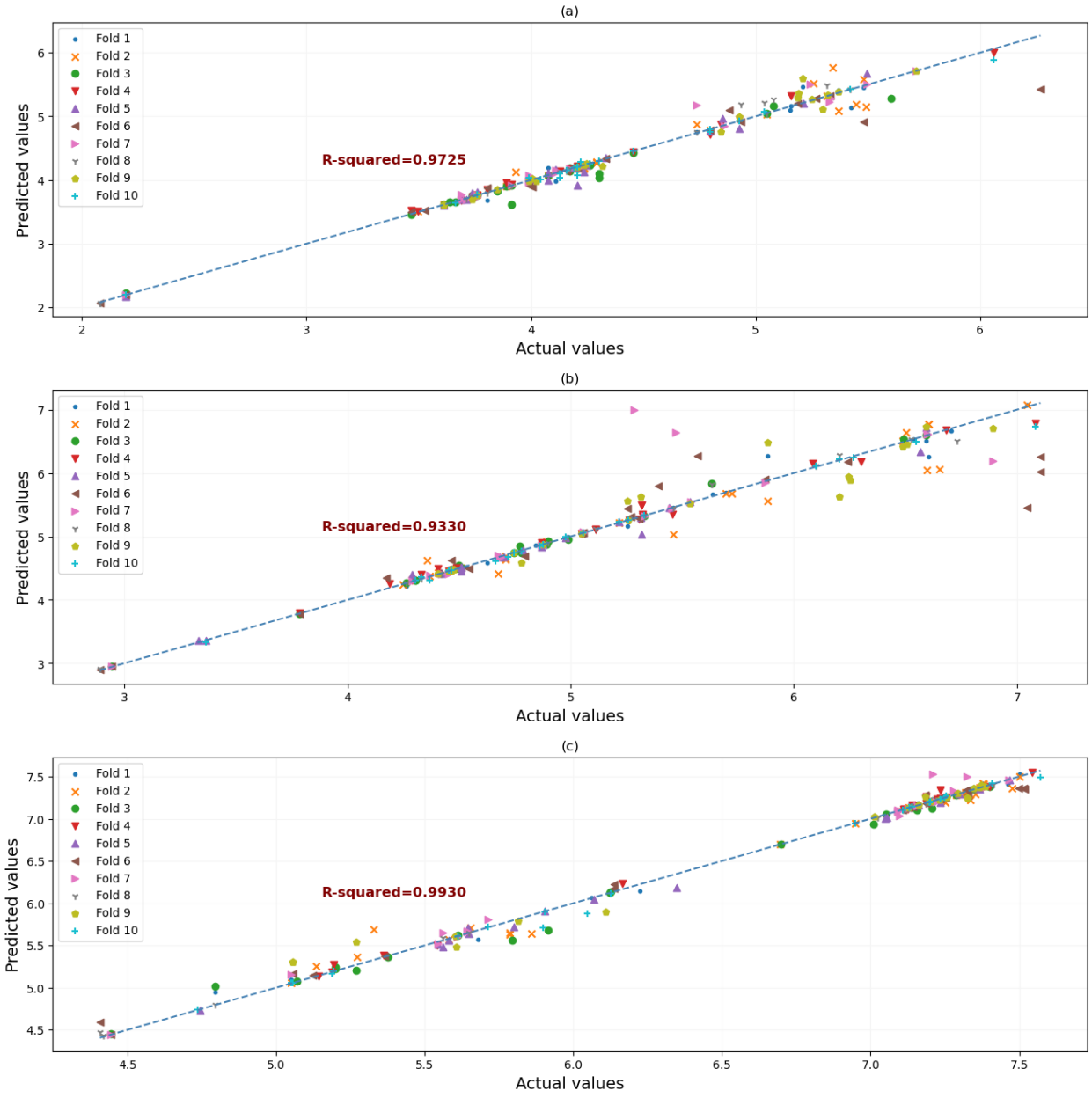

Figure 5. Actual vs. Predicted Values (10-fold cross validation) for VM: (a) Distance for 9% CO_2, (b) Distance for 4% CO_2, and (c) Distance for 1% CO_2.

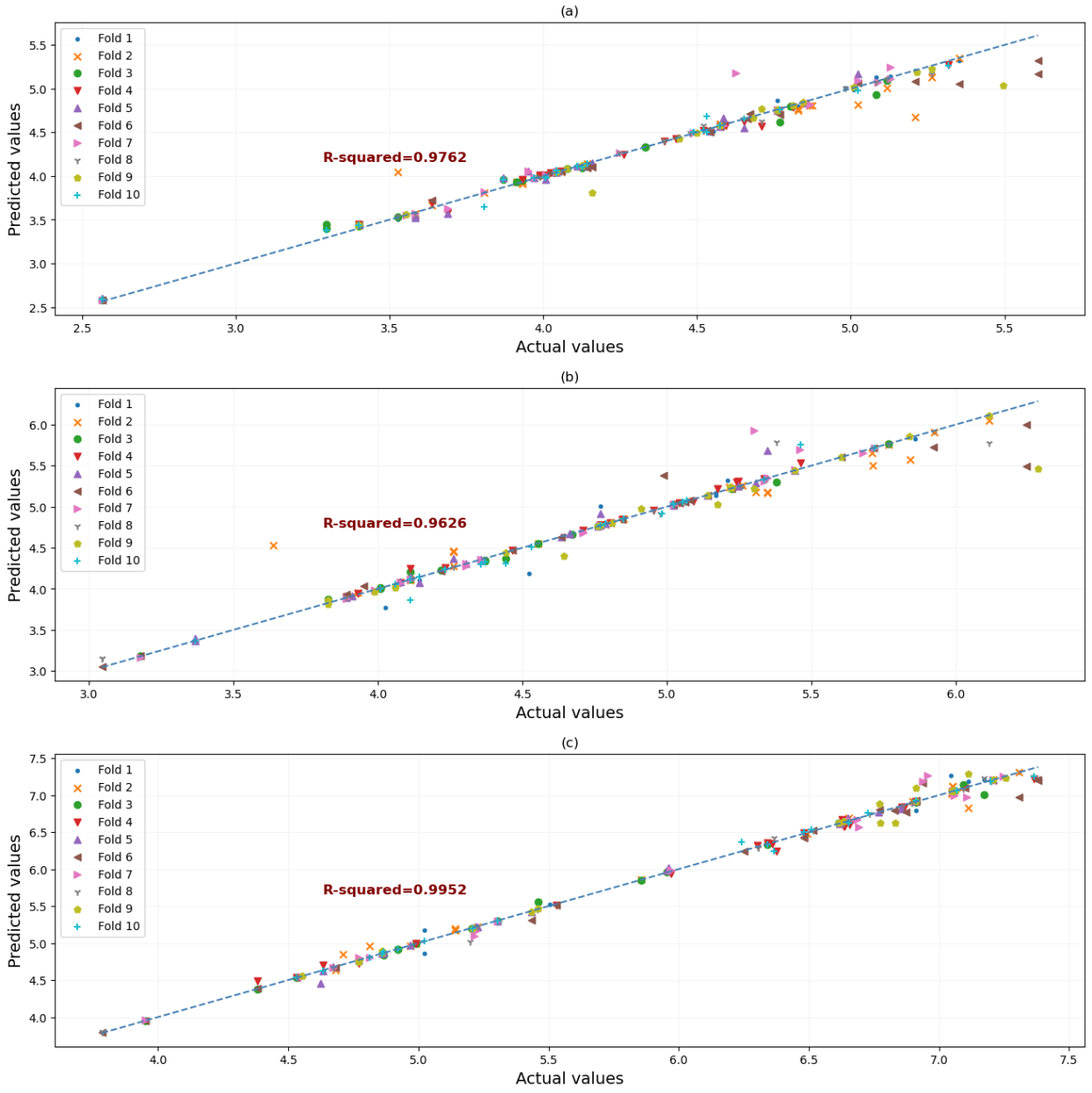

Figure 6. Actual vs. Predicted Values (10-fold cross validation) for VB: (a) Distance for 9% CO_2, (b) Distance for 4% CO_2, and (c) Distance for 1% CO_2.