Effect of Selection Procedures of High-Fidelity Data in Multi-Fidelity Surrogate Modeling

A. Özden*, A. Procacci*, R. Malpica Galassi*, F. Contino, A. Parente***

aysu.ozden@ulb.be

*Aero-Thermo-Mechanics Lab., Université Libre de Bruxelles, Avenue F. D. Roosevelt, 1050, Brussels, Belgium

** UC Louvain, Institute of Mechanics, Materials, and Civil Engineering, Louvain-la-Neuve, Belgium

Abstract

This study introduces a multi-fidelity digital twin for a combustion furnace that operates under MILD conditions. The digital twin aims to accurately predict the 3-dimensional reacting flow field at different operating conditions while keeping computational costs low. The training data comprises RANS simulations with detailed chemistry that cover the design space, which includes the H2 mole fraction (0-100%), equivalence ratio (0.7-1), and injector diameter (16, 20, and 25 mm). To reduce the costs of the training, lower fidelity 2-dimensional simulations replace some of the 3-dimensional simulations. The multi-fidelity model development strategy involves data compression using Principal Component Analysis, data fusion of the different fidelity inputs using Manifold Alignment, and interpolation of the latent variables with CoKriging. The trained model exhibits predictions that depend on the number of high-fidelity data replaced by lower-fidelity data. Therefore, an incremental sampling strategy is proposed to determine the minimal amount and location in design space of the high-fidelity data to fulfill accuracy requirements within a computational budget. Specific subsets of design conditions are found to be a better fit than others, and promising results are obtained in terms of temperature and species accuracy.

Introduction

Despite the increase in energy-related CO2 emissions due to persistent demand for fossil fuels, the combustion industry is exploring solutions like Moderate and Intense Low-oxygen Dilution (MILD) [1] combustion to meet emission targets. However, designing and understanding these systems require accurate and affordable tools to predict system response across a broad range of operating conditions. Reduced Order Models (ROMs) [2] have emerged as an effective tool, but for complex systems like combustion, a single fidelity ROM may not be practical. To address this, the Multi-fidelity Reduced Order Model (MF-ROM) combines sparse high-fidelity data with dense low-fidelity data to achieve acceptable accuracy within a fixed computational budget. This approach has been applied successfully in various areas over the past decade, such as in modeling an RAE 2822 airfoil [3]. This paper aims to develop a

multi-fidelity surrogate model of a combustion furnace in the MILD combustion regime that can predict its performance with acceptable accuracy using H_2 mole fraction, equivalence ratio, and injector diameter as design parameters. The developed surrogate model shows promising results, and further work is needed to fine-tune the model and improve its accuracy.

Methodology

Principal Component Analysis

The Principal Component Analysis (PCA) is a method of reducing the dimensionality of a dataset $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{nxp}}$, which contains n observations of p features, assuming that there are a large number of interrelated variables in the dataset [4]. In this study, CFD simulation results are used to construct the datasets. In this dataset, n observations are the number of numerical simulations while p represents the number of the number of grid points times the number of features, such as temperature and select chemical species mass fractions. In combustion problems, p is typically much larger than n. To perform PCA, $\mathbf{Z}$ of size ($n \times k$) and a matrix $\mathbf{\Phi}$ of size ($p \times k$) are computed, where $k \ll p$, by solving for the eigenvectors of the covariance matrix $\boldsymbol{C} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$. The columns of $\mathbf{\Phi}$ are known as the PCA modes, or eigenflames [5], and the matrix $\mathbf{Z}$ is called the PCA coefficients matrix. The lower dimension k is determined by the Relative Information content (RIC) value [6], which is interpreted as the total variance captured by the PCA. A threshold value of 99.9% or higher is commonly chosen. Since the eigenvectors are orthogonal and the principal components are linear combinations of the original variables, the high-fidelity response ($\mathbf{X}_{\mathbf{rec}}$) of an undiscovered point ($\mathbf{z}_{\mathbf{j}}$) in design space can be calculated as $\boldsymbol{X}_{\boldsymbol{rec}} = \mathbf{\Phi}\boldsymbol{z}_{\boldsymbol{j}}$.

Manifold Alignment

A multi-fidelity surrogate model combines information from datasets with different dimensions and topology to solve the same physical problem. To transfer information between these datasets, manifold alignment with the assumption of these datasets shares a common low-dimensional subspace is used [7]. Procrustes manifold alignment is a specific method that utilizes PCA to align the low-dimensional representations of these datasets, allowing for optimal alignment between them. This technique has been successfully applied in various fields, such as protein alignment [8] and image matching [9]. In this study, the Procrustes manifold alignment technique is used, which involves using PCA as the first step. Procrustes analysis is employed to remove translational, rotational, and scaling components from one manifold to achieve an optimal alignment with the other [10]. The high-dimensional and low-dimensional datasets are identified as $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{nxp}}$ and $\boldsymbol{Y} \in \mathbb{R}^{\boldsymbol{mxq}}$, respectively. In this case $n<m$ and the low-fidelity dataset involves a linked subset $\boldsymbol{Y}_{\boldsymbol{L}} \in \mathbb{R}^{\boldsymbol{nxq}}$, which shares the same input design parameters as the high-fidelity dataset. As the first step of the method, both outputs are subjected to PCA

for dimensionality reduction, and PCA modes $\mathbf{\Phi} \in \mathbb{R}^{pxk}$ and $\mathbf{\Psi} \in \mathbb{R}^{qxk}$, with PCA coefficients $\boldsymbol{Z} \in \mathbb{R}^{nxk}$ and $\boldsymbol{W} \in \mathbb{R}^{mxk}$, are obtained for high- and low-fidelity inputs, respectively. The objective of this transformation is to find the transformed manifold $\mathbf{T}$ that minimizes $\|\mathbf{Z}-\mathbf{T_L}\|$, and this orthogonal Procrustes manifold is solved using the procedure of Wang and Mahadevan [8]. The first step is to ensure that the linked datasets $\mathbf{Z}$ and $\mathbf{W_L}$ both have their centroids at the origin. Then, the scaling factor (s) and rotation matrix ($\mathbf{Q}$) can be obtained by computing the Singular Value Decomposition (SVD) $\boldsymbol{U\Sigma V^T} = \boldsymbol{W_L Z^T}$, where $U \in \mathbb{R}^{kxk}$, $V \in \mathbb{R}^{kxk}$, and $\Sigma \in \mathbb{R}^{kxk}$ are the left and right singular vectors and the diagonal matrix containing the singular values, respectively. Finally, after computing the scaling factor $s = \frac{tr(\Sigma)}{tr(W_L W_L^T)}$ and rotating matrix $Q = VU^T$, the transformed matrix, which is the low-fidelity approximation of $\mathbf{Z}$, can be obtained as $\boldsymbol{T} = \boldsymbol{sQW}$.

Regression Model: CoKriging

The Gaussian Process Regression (GPR) [11] is commonly used as an interpolation model in statistics and machine learning to predict output for undiscovered design points. When combining high- and low-fidelity data, a multi-fidelity regression model like CoKriging is necessary [12]. The CoKriging regression model used in this study follows the auto-regressive model of Kennedy and O'Hagan [12], which assumes that low-fidelity data cannot provide any additional insights if a high-fidelity counterpart is available [13]. The process of generating a CoKriging model involves constructing two Kriging models in sequence. First, a Kriging model based on low-fidelity data is constructed. Then, a second Kriging model is created to capture the discrepancies between the high-fidelity and low-fidelity data, expressed as $\hat{Y}_H = \rho Y_L + \delta$, where $\hat{Y}_H$ represents the high-fidelity approximation, ρ is a constant scaling factor, and δ is the difference between the high- and low-fidelity models.

Dataset

This study uses 3D and 2D computational fluid dynamics (CFD) simulations of a semi-industrial MILD combustion furnace with a nominal power of 20 kW. The simulations are conducted using ANSYS Fluent 19.1 software [14]. The standard k-ε turbulence model is used in combination with the PaSR [15] model for turbulence-chemistry interactions, and the kinetic scheme employed is the Kee mechanism with 17 species and 58 reactions. Radiation is modeled using the discrete ordinate (DO) method with the weighted-sum-of-gray-gases (WSGG) model. The high- and low-fidelity simulations have grid sizes of 216360 and 28683, respectively. The fuel composition, equivalence ratio, and injector diameter collectively define the design space with the ranges of 0-100% for H2 mole fraction, 0.7-1 for equivalence ratio (φ), and 16-20-25mm for the air injector diameter. The design of experiments (DoE) is generated using latin hypercube sampling (LHS) method, as previously performed by Aversano et al. [16], with a total of 45 simulations performed and 4 test cases

excluded for testing purposes. The simulations are used to train the regression model, with the features of interest being temperature and major (CH_4, H_2, O_2, H_2O, OH) /minor species (CO and OH). The design parameters for the test simulations are listed in Table 1.

Table 1. Design conditions for test cases from Aversano et al. [16]

Simulation No	Air Inj. Dia. (mm)	H_2 (%)	ER (-)
1	16	60	0.93
2	20	50	0.83
3	20	90	0.74
4	25	65	0.91

The crucial decision is to determine the selection and quantity of simulations for constructing the high-fidelity dataset. For an initial comparison, 10 high-fidelity simulations are utilized, chosen based on the leave-one-out analysis conducted by Aversano et al. [22]. As previously stated, the low-fidelity dataset is separated into 10 connected data points with identical design parameters to the 10 high-fidelity solutions, which are required for the manifold alignment process, as well as 31 additional unconnected data points.

Results

The Normalized Root Mean Square Error (NRMSE) is a frequently used metric for assessing the accuracy of predictions compared to the results of CFD simulations for the test cases. NRMSE is calculated by dividing the Root Mean Square Error (RMSE) by the mean of the observations:

$$NRMSE = \frac{1}{\overline{X_{obs}}}\sqrt{\frac{\sum_{i=1}^{N}\left(X_{obs,i}-X_{pred,i}\right)^2}{n}} \quad (1)$$

Figure 1 displays a comparison of the temperature and OH fields, respectively, for test simulation #4 between the CFD and multi-fidelity ROM. The temperature field prediction demonstrates that the MF-ROM can accurately predict the temperature distribution, closely resembling the original data. In terms of the OH field, although there is room for improvement of low-fidelity, the overall species distribution appears satisfactory.

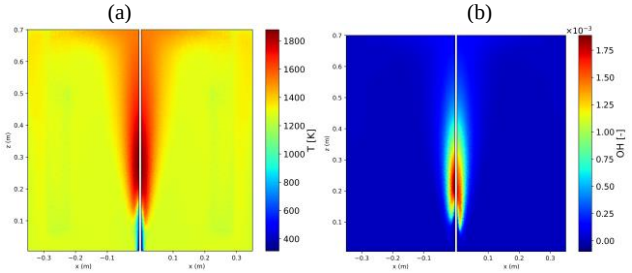

Figure 1. Temperature (a) and OH (b) field comparison between the CFD simulation *(left)*, and the multi-fidelity ROM prediction *(right)*, case #4.

The current study has highlighted the sensitivity of the prediction accuracy to the number and distribution of high-fidelity simulations in the design space. To understand the selection of simulations, an incremental sampling decision-making strategy was developed, which involves adding high-fidelity samples where the trained model exhibits the largest uncertainty. The effect of additional samples is evaluated using 10-fold cross-validation on the train/test data split, and the results are presented in Figure 2, showing that error metrics decrease almost monotonically with increasing high-fidelity samples, confirming that the incremental sampling strategy is well-posed. A saturation point is also observed after the addition of ~20 high-fidelity samples. Additionally, when certain cases are selected as test cases, NRMSE values can be amplified by up to 200%, suggesting that these cases should be included in the training dataset.

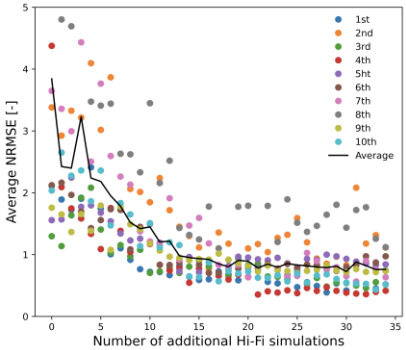

Figure 2. Impact of additional high-fidelity samples for 10-fold cross validation

Conclusion

This study aimed to develop a multi-fidelity surrogate model for predicting the three-dimensional spatial fields of a MILD combustion furnace based on the design parameters of air injector diameter, H_2 mole fraction, and equivalence ratio. The approach exploits Principal Component Analysis, Procrustes Manifold Alignment, and CoKriging to project data in a shared latent space and interpolate the latent variables over the whole design space. 45 high-fidelity and low-fidelity simulations are used to build the model. A model trained with 10 high-fidelity and 41 low-fidelity simulations exhibited prediction errors below 10% for temperature and H_2O, but errors were higher for other species, especially CH_4, H_2, OH, and CO. However, temperature and OH reconstructed fields showed good qualitative agreement with the CFD counterpart. Incremental sampling was found to be crucial in determining the optimal number of simulations and design locations in the DoE of high-fidelity simulations to meet the accuracy requirement for a given computational budget. An uncertainty-based selection method indicated that specific subsets of design conditions are a better fit than others for the high-fidelity dataset to improve prediction accuracy. Promising results were obtained for the determination of an optimal number of simulations, but further investigation is needed to fine-tune CoKriging hyper-parameters and formalize the results of the sampling strategy.

References

[1] Cavaliere, A., De Joannon, M., "Mild combustion", *Progress in Energy and Combustion Science*, 30 (4):329-366 (2004)

[2] Lucia, D.J., Beran, P.S., Silva, W.A., "Reduced order modeling: new approaches for computational physics", *Progress in Aerospace Sciences*, 40 (1-2):51-117 (2004)

[3] Han, Z.H., Zimmermann, Görtz, S., "Alternative cokriging method for variable-fidelity surrogate modeling", *AIAA journal*, 50 (5):1205:1210 (2012)

[4] Jolliffe, T.J., *Principal component analysis for special types of data*, Springer, 2002.

[5] Aversano, G., Bellemans, A., Li, Z., Coussement, A., Gicquel, O, Parente, A., "Application of reduced-order models based on pca & kriging for the development of digital twins of reacting flow applications", *Computers & chemical engineering* 121:422-441 (2019).

[6] Pinnau, R., *Model order reduction: theory, research aspects and applications*, Springer, 2008, pp. 95–109.

[7] Wang, C., Mahadevan, S., "A general framework for manifold alignment", *2009 AAAI Fall Symposium Series* (2009)

[8] Wang, C., Mahadevan, S., "Manifold alignment using procrustes analysis" *Proceedings of the 25th international conference on Machine learning*, pp. 1120-1127 (2008)

[9] Ham, J., Lee, D., Saul, L., "Semisupervised alignment of manifolds" *International Workshop on Artificial Intelligence and Statistics*, pp. 120-127. (2005)

[10] Luo, B., Hancock, E.R., "Feature matching with Procrustes alignment and graph editing", pp. 72-76, (1999)

[11] Williams CK, Rasmussen CE., "Gaussian processes for machine learning", *MA: MIT press Cambridge*, vol:2 (2006)

[12] Kennedy, M.C., O'Hagan, A., "Predicting the output from a complex computer code when fast approximations are available", *Biometrika*, 87(1):1-13 (2000)

[13] Forrester, A.I., Sóbester, A., Keane, A.J., "Multi-fidelity optimization via surrogate modelling", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088): 3251-3269 (2007)

[14] M. Ferrarotti, M. FÅNurst, E. Cresci, W. De Paepe, and A. Parente, "Key modeling aspects in the simulation of a quasi-industrial 20 kw moderate or intense low-oxygen dilution combustion chamber", *Energy & fuels*, 32(10): 10228-10241 (2018)

[15] Chomiak, J., "Combustion a study in theory, fact and application", (1990)

[16] Aversano, G., Ferrarotti, M., Parente, A., "Digital twin of a combustion furnace operating in flameless conditions: reduced-order model development from CFD simulations", *Proceedings of the Combustion Institute*, *38*(4):5373-5381 (2021)