# Data-Driven Stochastic Turbulence Generation via Physics-Informed Field Inversion Machine Learning

**Levent Ugur[1] and Beckett Y. Zhou[1]**

[1] Daniel Guggenheim Aerospace Engineering Department, Georgia Institute of Technology, Atlanta, 30332-0150, GA, USA

**Keywords:** Synthetic turbulence, Stochastic Noise Generation, data assimilation, physics-informed loss, sparse observations

**Abstract.** Stochastic Noise Generation (SNG) is a low-fidelity method enabling stochastic turbulent velocity field generation based on a given steady flow solution. In the previous study by the authors, the performance of SNG is enhanced by employing the Field Inversion Machine Learning (FIML) technique to assimilate the available high-fidelity data. Although the outcomes of this study were highly promising, the FIML-SNG framework relies on dense observations, which limits its ability to leverage sparse observations, such as experimental data cases. To address this limitation, we propose a Physics-Informed Field Inversion (PIFI) modification, which employs the residual of the convection equation as the physical loss by aiming to obtain dense gradient information during the inverse optimization stage. The proposed PIFI approach is applied to 15 different jet flow realizations by leveraging the data recorded at only six probe locations. By using the optimal spatial and modal correction parameters obtained from this stage, two distinct machine learning models are trained for later use in predictive mode. The trained models are then tested on both seen and unseen cases. The results of the PI-FIML-SNG framework indicate that the proposed PIFI modification improves the reconstruction performance for sparse observations.

## 1. Introduction

Despite time-averaged flow statistics being sufficient for most of the aerodynamic shape design studies, time-dependent statistics are still required in some key design considerations such as reducing the aerodynamic noise. However, given the high computational cost of unsteady CFD simulations, performing efficient design cycles with high-fidelity solvers requires a great computational resource. Synthetic turbulence generation can stand as an enabler for that purpose to model the time-dependent flow field based on a steady flow solution. One of the methods to generate a turbulent velocity field is Stochastic Noise Generation (SNG) [1, 2], primarily developed to predict turbulent flow statistics necessary for farfield aeroacoustic noise predictions. SNG takes pointwise inputs from a RANS solution and predicts pointwise time-dependent turbulence velocity components by using a Fourier series approximation to model the turbulent flow as a summation of modes with varying frequencies.

Although SNG offers promising performance for use in aeroacoustic noise prediction frameworks [3] and is even employed in some shape design studies [4], its capability in predicting highly accurate turbulent velocity fields is still limited given the low-fidelity nature of the model. The previous study of the authors [5] suggests enhancing the prediction performance of the SNG model by employing the Field Inversion Machine Learning (FIML) [6] technique, which is predominantly applied to correct turbulence closure models of RANS solvers. FIML typically suggests correcting the free parameters of a low-fidelity model spatially by predicting correction coefficients from machine learning (ML) models trained on the assimilated knowledge from high-fidelity observations after a field inversion stage. In our previous study, a modal correction set was also introduced, in addition to the spatial corrections, to correct not only the time-averaged statistics but also the time-dependent statistics of the generated turbulent flow field. The results of the FIML-SNG framework indicate the potential of such a synthetic turbulence generation model enhanced with high-fidelity observations.

While FIML is an advanced data-driven method for such a problem, its solely data-driven nature has some limitations. The Bayesian inversion-based field inversion stage of the framework requires dense correlation matrices to enable field inversion with sparse observations. However, for the cases with limited prior knowledge, calculating useful correlation matrices may not be possible, and this can impact the success of the sparse data assimilation, especially for the cases using pointwise low-fidelity models such as SNG.

In this study, we employ a Physics-Informed Field Inversion (PIFI) approach to extend our previous study to enable data assimilation with sparse observations. This modified field inversion stage suggests embedding the physical background of the problem to the inverse optimization stage in the form of a physical loss, inspired by the Physics-Informed Neural Networks [7]. To correct the SNG model, we employ the 3D convection equation residual as the physical loss term, which is also in line with the study of Billson et al. [3] highlighting the necessity of explicit filtering of the SNG predictions to obtain convection-aware results. Our convection loss addition can also mitigate this weakness of the SNG.

To test the proposed PI-FIML-SNG approach, 15 jet flow cases, sharing the same round jet geometry but varying input profiles, are employed. Their high-fidelity data are obtained by LES analyses, with observations at only six probe locations used during training, akin to an experimental setup. Spatial and modal correction parameters are optimized using data and the convection loss to correct both time-averaged and time-dependent statistics of the SNG predictions. The machine learning models are then trained using the extracted knowledge from these cases and tested on both seen and unseen cases. The proposed PI-FIML-SNG approach is summarized in Figure 1.

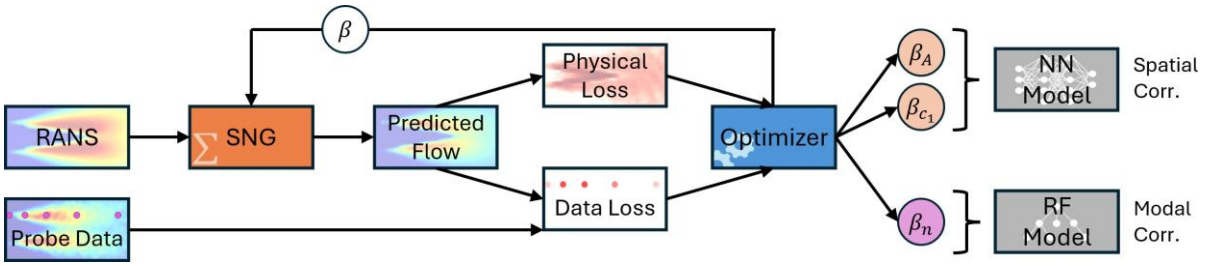


**Figure 1.** Proposed PI-FIML-SNG Framework.

## 2. Methodology

### *2.1. Stochastic Noise Generation*

Stochastic Noise Generation (SNG) [1, 2] is a form of synthetic turbulence generation method that is primarily developed to predict unsteady turbulent flow statistics to be used in aeroacoustics solver frameworks cost-effectively. The model explained in this section is based on our previous study [5] which has some modifications compared to the earlier studies.

$$u_t(x,t) = 2 \sum_{n=1}^{N} \hat{u}_n \cos(k_n x + \psi\psi_n + \omega_n t)\, \boldsymbol{\sigma}_n \tag{1}$$

Equation (1) shows the Fourier series approximation used by the SNG to model the turbulent velocity, $u_t$. In this equation, $N$ stands for the number of modes, whose corresponding frequencies are determined by the $\omega_n$ term. A modification that distinguishes our baseline model used in this study from the previous one is the calculation of this frequency term, such that it becomes a function of turbulent kinetic energy dissipation rate, $\epsilon$ obtained from RANS solution as given in Equation (2).

$$\omega_n = \sqrt{\epsilon \|k_n\|} \tag{2}$$

In these equations, wave vector, $k_n$; phase angle, $\psi\psi_n$; and unit vector, $\boldsymbol{\sigma}_n$ are random variables whose generation is discussed in [2]. Another term that is important as much as the frequency term is the magnitude of the velocity of each mode, $\hat{u}_n$. This term is modeled as a function of the wavenumber as $\hat{u}_n = \sqrt{E(\|k_n\|)\ \Delta\|k_n\|}$, where the function $E$ is the von Karman-Pao isotropic turbulence formula given in Equation (3).

$$E(k) = \frac{2A}{3}\frac{K}{k_e}\left(\frac{k}{k_e}\right)^4 \exp\left(-2\left(\frac{k}{k_\eta}\right)^2\right)\left(1+\left(\frac{k}{k_e}\right)^2\right)^{-17/6} \quad (3)$$

Equation (3) takes turbulent kinetic energy dissipation rate, $\epsilon$ and kinematic viscosity, $\nu\nu$ obtained from the RANS solution as the inputs to calculate $k_\eta = \epsilon^{1/4}\nu^{-3/4}$. Similarly, $k_e$ can be calculated by using the total turbulent kinetic energy, $K$ predicted by RANS such that $k_e = 0.747/L_T$ where $L_T = c_1 u'/\epsilon$ and $u' = \sqrt{2K/3}$. Further details of the method can be found in the relevant references [1-5].

There are two free parameters in the equations of the SNG model: $A$ and $c_1$. These constants are empirically calculated as 1.453 and 1.0, respectively, in [2]. However, Billson et al. [3] discussed the necessity of tuning these parameters for different problems. Our ultimate aim in this study is to correct these parameters with a high degree of freedom to obtain more accurate predictions from the SNG model by assimilating the available high-fidelity data. To perform this correction, two spatially varying correction parameter sets, $\beta_A$ and $\beta_{c_1}$, are employed, which are applied as multipliers to the corresponding free parameters. Additionally, as we desire to correct not only the time-averaged statistics but also the time-dependent ones, we introduce an additional modal correction set, $\beta_n$, applied to the $\sqrt{k_n}$ term, again as a multiplier, to increase the control over the temporal behavior of the generated turbulent field.

### *2.2. Physics-Informed Field Inversion*

Field Inversion Machine Learning (FIML) approach, originally introduced in [6], is an advanced data assimilation framework consisting of two stages. The first stage is field inversion which aims to decrease the discrepancy between the low-fidelity model predictions and the high-fidelity observations through a Bayesian inversion to find the optimal model parameters. However, it can be challenging to calculate useful correlation matrices for sparse observation cases, and hence, simplifying the inversion problem to ridge regression can be necessary. Under this simplification, the objective components responsible for the data losses can be chosen as Mean Squared Error (MSE) between the observations and predictions calculated for $N_p$ probe locations, as used in Equations (4) and (5). In this study, we considered two different quantities for the data loss calculation: diagonal Renolds Stress Tensor (RST) components for time-averaged turbulent flow statistics, and Power Spectral Density (PSD) of the velocity recordings at the probe locations for the time-dependent statistics.

$$J_{\text{RST}} = \frac{1}{3N_p}\sum_{p=1}^{N_p}\sum_{jj=1}^{3}\left(\left(\overline{u_y u_y}\right)^p_{\text{SNG}} - \left(\overline{u_y u_y}\right)^p_{\text{LES}}\right)^2 \quad (4)$$

$$J_{\text{PSD}} = \frac{1}{3N_p}\sum_{p=1}^{N_p}\sum_{jj=1}^{3}\left(\log\left(\text{PSD}(u'_{jj})^p_{\text{SNG}} - \log\left(\text{PSD}(u'_{jj})\right)^p_{\text{LES}}\right)^2 \quad (5)$$

Although we mentioned that assuming uncorrelated quantities can be sometimes necessary, such an assumption is not essentially correct for physical systems and can dramatically decrease the recovery performance of the inversion. The key contribution of this study is modifying the FIML-SNG framework with a Physics-Informed Field Inversion (PIFI) approach which suggests a physical loss term addition during the field inversion stage in addition to the data loss. In this study, the residual of the convection equation is employed as the physical loss as shown in Equation (6), averaged over all grid points used in the inversion, $N_{gg}$. This loss enables us to create a connection between independent solution points of the SNG model and provides dense gradient information for every correction parameter during the inverse optimization stage.

$$J_{\text{phys}} = \frac{1}{3N_{gg}}\sum_{gg=1}^{N_g}\sum_{jj=1}^{3}\left(\frac{\partial(\bar{\rho}u'_{jj})}{\partial t} + \frac{\partial(\bar{\rho}\tilde{u}_k u'_{jj})}{\partial x_k}\right)^2 \quad (6)$$

The defined objective components are combined to form the total objective to be minimized during the inverse optimization stage by using a weighted summation. After a search study, the loss weights shown in Equation (7) are determined as $\alpha_{\text{RST}} = 1.0, \alpha_{\text{PSD}} = \alpha_{\text{phys}} = 0.05$ and $\alpha_{\text{reg}} = 0.001$. Each objective

component is normalized by using its baseline value and the overall objective is scaled with $1.0 \times 10^6$ as this setting offers a better convergence. The optimal $\beta$ values are obtained by employing the L-BFGS-B implementation of the SciPy package to minimize the combined objective provided.

$$J = \alpha_{\mathrm{RST}} J_{\mathrm{RST}} + \alpha_{\mathrm{PSD}} J_{\mathrm{PSD}} + \alpha_{\mathrm{phys}} J_{\mathrm{phys}} + \alpha_{\mathrm{reg}} J_{\mathrm{reg}} \quad (7)$$

An important note on the implementation of the framework is that as the number of correction parameters scales with the number of grid points to be used during the inversion, the problem can be considered a large-scale optimization problem requiring an efficient gradient calculation for the correction parameters. For that purpose, the SNG model is implemented by using the JAX library which enables usage of GPU during the linear algebra operations and provides adjoint sensitivities thanks to its automatic differentiation features.

### *2.3. Training of the Machine Learning Models*

Once the optimal correction parameters are obtained after the PIFI stage for every realization, the datasets to be used in the second stage of the FIML framework are prepared. In this stage, machine learning models are trained by leveraging the optimal corrections obtained in the PIFI stage so that the trained models can be used in predictive mode to predict appropriate correction parameters for any given case. In our study, two distinct machine learning models are trained for the spatial and the modal corrections.

*Neural Network Model for Spatial Corrections.* The machine learning model training for spatial corrections is more straightforward and in line with the traditional FIML applications. The features are chosen as the pointwise mean flow properties whereas the labels are optimal $\beta_A$ and $\beta_{c_1}$ values. To create a mapping to this regression problem, a neural network model with 3 hidden layers, 30 nodes each is employed. The training is performed for 1000 epochs by using the same settings provided in [5].

*Random Forest Model for Modal Corrections.* An unconventional approach used in the FIML-SNG framework is utilizing the modal corrections, while the FIML technique typically suggests using only spatial corrections to create a mapping between pointwise inputs of the low-fidelity model and the optimal corrections. However, such a natural mapping does not exist for the modal corrections if there are no time-dependent or modal inputs of the low-fidelity model. To address this challenge, we proposed in our previous study [5] using the distribution of turbulence closure model outputs, turbulent kinetic energy and its dissipation rate, along the lipline of the nozzle since these outputs are the best candidates that can represent some information about the unsteady behavior of the flow. In this study, we use the same approach with the difference being that we consider only the peak points of the optimal modal correction distributions as the training samples. The main motivation behind this choice is that these peaks are not ordinary outliners, but dominant correction parameters applied to the SNG model, and they should be predicted accurately. A random forest model is trained to fit a function to the proposed mapping using the extracted samples.

## 3. Results & Discussion

By employing the described methodology, the Physics-Informed Field Inversion is performed for 15 different jet flow cases. These cases share the same geometry but varying inlet conditions as described in [8]. The observations are obtained from LES results by considering only six probe locations along the lipline of the nozzle located at $x/D = [0.5, 1.0, 2.5, 5.0, 10.0, 15.0]$. To mimic an experimental setup, it is assumed that both RST and PSD statistics are known only at these probe locations in the solutions space, and the RST statistics of the whole space are only utilized to test the inversion results. RANS with $k$-$\omega$ SST turbulence closure model and LES solutions are obtained by employing SU2 solvers, and then interpolated onto a coarser, structured grid with the size of (61, 15, 15) to be used during the inverse optimization stage. After the PIFI stage is completed for all realizations, the datasets with optimal spatial and modal corrections are generated and the corresponding machine learning models are trained using these datasets.

*Seen Case Results.* To assess the success of the PIFI approach in this sparse data optimization case, the optimized results are first presented for a seen case, one of the 15 flow cases used in the training of the ML models. The RST statistics of SNG models using different correction sets are compared in Figure 2 in the form of the average of root squared errors of all three diagonal components at the meridian plane. This comparison shows that the SNG model using the optimal correction sets obtained from the PIFI approach decreases the discrepancy between the model predictions and high-fidelity LES data compared with the baseline SNG model, resulting in an increase in the recovery score from 0.68 to 0.91. On the other hand, the SNG model using the optimal corrections obtained from the only data-driven FI approach overfits the data at the probe locations and increases the error for the rest of the solution space. Note that this global effect is due to the modal corrections having the same impact across all grid points. These modal corrections are efficiently tuned by both PIFI and conventional FI such that they provide closer PSD trends to the LES solution than the baseline model, as can be seen in Figure 3.

The ML-enhanced SNG results, using the corrections predicted by the trained ML models, are also compared with the other results to evaluate their prediction capability. For the PSD statistics, we see that the ML-enhanced SNG model again provides better trends than the baseline SNG and the results mostly match with the optimized SNG predictions. For this sample case, the RST recovery score of the ML-enhanced SNG model takes 0.73 and stands between the optimized model with PIFI and the baseline model, with a closer result to the baseline. Although a score in this interval is already expected, we should note that this is not always the case for every realization. In some seen cases, the ML-enhanced results can be outperformed by the baseline model in terms of RST statistics, although they still improve the PSD statistics, whereas in some other cases, the ML-enhanced results can even surpass the optimized SNG model with PIFI. None of these results are expected outcomes, and our investigations indicate that this uncertainty is mainly due to the poor recovery performance of the trained random forest model for modal corrections. We also observed that the RST recovery accuracy is highly sensitive to these corrections, particularly those affecting low-frequency terms, and their inaccuracies significantly impact the results. This shows that further effort is necessary to improve the modal correction predictions.

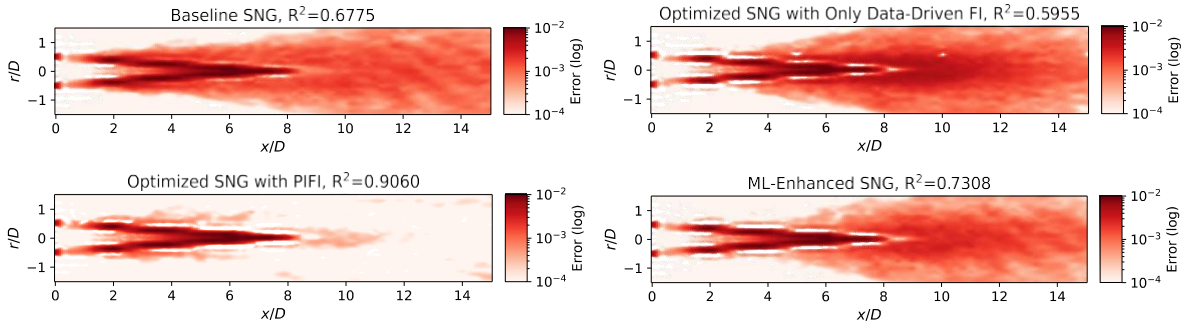


**Figure 2.** Mean of the root squared error distribution for the diagonal RST components between LES data and SNG model predictions at the meridian plane for the seen case.

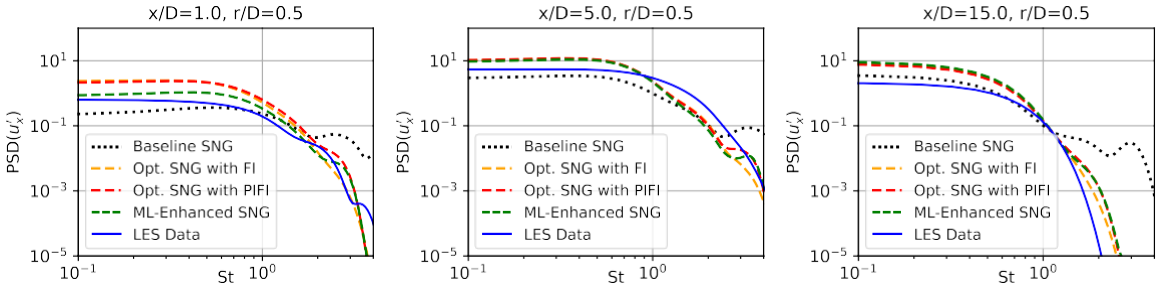


**Figure 3.** Comparison of streamwise velocity PSD statistics at different probe locations for the seen case.

*Unseen Case Results.* The same trained ML models are also tested on an unseen case, which is crucial for commenting on the generalizability of the knowledge obtained from the training cases. It is evident that the ML-enhanced SNG results outperform the baseline model in terms of both RST and PSD

statistics, as shown in Figures 4 and 5, respectively, showcasing a notable improvement in the generated turbulent flow field, especially for downstream locations. It should be noted that the high uncertainty of the ML-enhanced SNG predictions, reported for the seen cases, can again be an issue for the unseen cases, which can increase or decrease the performance relative to this case. Nevertheless, the success of this test case demonstrates that the PI-FIML-SNG framework can be a promising approach for the turbulent flow field reconstruction for sparse observation cases.

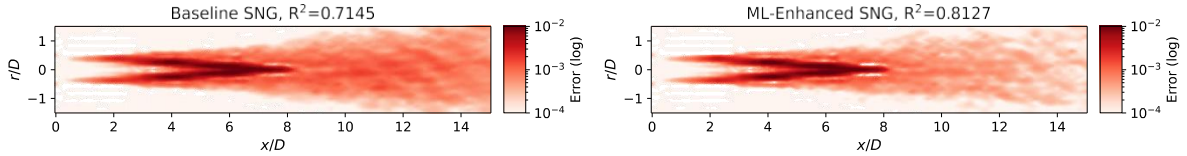


**Figure 4.** Mean of the root squared error distribution for the diagonal RST components between LES data and SNG model predictions at the meridian plane for the unseen case.

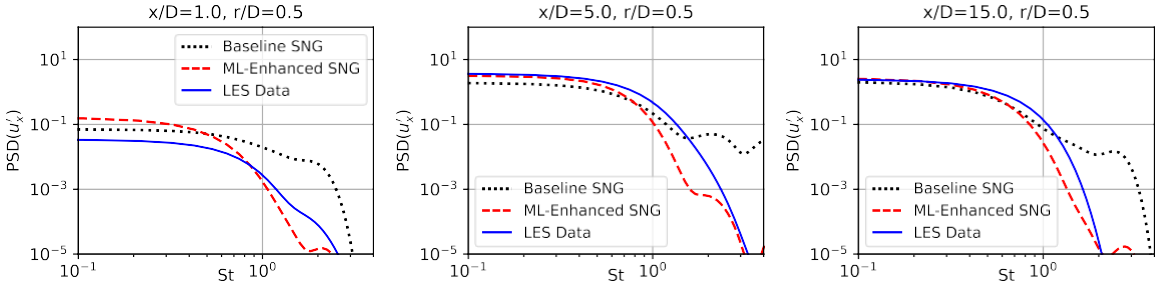


**Figure 5.** Comparison of streamwise velocity PSD statistics at different probe locations for the unseen case.

## 4. Conclusion

The field inversion results obtained in this study indicate that the PIFI approach can be a beneficial modification in the FIML framework to obtain optimal correction parameters in sparse observation cases to enhance the accuracy of both time-averaged and time-dependent statistics of the turbulent flow field generated by the SNG model. However, to be able to use this assimilated knowledge in other cases to model realistic synthetic turbulence, accurate machine learning model training is also necessary. Although the trained ML model for spatial corrections performs well in predicting the optimal corrections in this study, it has been seen that the performance of the random forest model, used for the modal correction predictions, can lead to uncertainties in the ML-enhanced SNG model predictions. Although the current effort in this paper on creating a mapping for the modal corrections can be a promising baseline, it should be further improved in future studies for more stable and accurate predictions of these coefficients.

## References

[1] W. Bechara, et al., 'Stochastic approach to noise modeling for free turbulent flows', AIAA J., vol. 32, no. 3, pp. 455-463, 1994. https://doi.org/10.2514/3.12008.
[2] C. Bailly, D. Juve, 'A stochastic approach to compute subsonic noise using linearized Euler's equations', 5th AIAA/CEAS Aeroacoustics Conference and Exhibit, 1999. https://doi.org/10.2514/6.1999-1872.
[3] M. Billson, L.-E. Eriksson, L. Davidson, 'Jet Noise Prediction Using Stochastic Turbulence Modeling', 9th AIAA/CEAS Aeroacoustics Conference and Exhibit, May 2003. https://doi.org/10.2514/6.2003-3282.
[4] B. Y. Zhou, et al., 'Adjoint-based broadband noise minimization using stochastic noise generation', 25th AIAA/CEAS Aeroacoustics Conference, 2019. https://doi.org/10.2514/6.2019-2697.
[5] L. Ugur, F. T. Kunz, B. Y. Zhou, 'Field Inversion Machine Learning Based Stochastic Noise Generation Model for Jet Noise Prediction', 30th AIAA/CEAS Aeroacoustics Conference, 2024. https://doi.org/10.2514/6.2024-3127.
[6] E. J. Parish, K. Duraisamy, 'A paradigm for data-driven predictive modeling using field inversion and machine learning', J. Comput. Phys., vol. 305, pp. 758-774, 2016. https://doi.org/10.1016/j.jcp.2015.11.012.
[7] M. Raissi, P. Perdikaris, G. E. Karniadakis, 'Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations', J. Comput. Phys., vol. 378, pp. 686-707, 2019. https://doi.org/10.1016/j.jcp.2018.10.045.
[8] A. Jivani, et al., 'Uncertainty quantification for a turbulent round jet using multifidelity Karhunen-Loève expansions', AIAA Scitech 2021 Forum, 2021. https://doi.org/10.2514/6.2021-1367.