Generative subgrid-scale modeling

Jiaxi Zhao[a], **Qianxiao Li**[b]

[a] *National University of Singapore,* jiaxi.zhao@u.nus.edu

[b] *National University of Singapore, qianxiao@nus.edu.sg*

* Presenting author

1. Introduction

The mismatch between the a-priori and a-posteriori error is ubiquitous in data-driven subgrid-scale (SGS) modeling, which is an important ingredient in large eddy simulations. The a-priori error refers to the error of the closure modeling, usually as a regression problem, from the dataset generated by the fine-grid simulation while the a-posteriori error represents the error arised when this data-driven SGS model is deployed to another simulation. In this work, we investigate the cause of this mismatch in depth and the contribution of this work is two-fold:

1. We explain the mismatch of "a-priori and a-posteriori dichotomy", by studying the dataset for SGS model training and identify two features: data imbalance and multi-valuedness.
2. We propose a generative modeling of the SGS stresses via a conditional Gaussian model to resolve the multi-valuedness and demonstrate improvement in the simulation of the KS equation.

2. A-priori error analysis for regression-based data-driven SGS modeling

In this paper, we use the Navier-Stokes (NS) and the Kuramoto-Sivashinsky (KS) equations as examples for our illustrations. The NS equations are given by:

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j}, \quad \frac{\partial u_i}{\partial x_i} = 0, \tag{1}$$

Consider a spatially-homogeneous filter G, e.g. Gaussian filter $G(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^3}} e^{-\|\mathbf{x}\|_2^2/2}$, convolved to the field variable $\mathbf{u}$ componentwise, i.e. $\widetilde{\mathbf{u}} = G * \mathbf{u}$, one can obtain the equation governing the filtered field variable $\widetilde{\mathbf{u}}$:

$$\frac{\partial \widetilde{u}_i}{\partial t} + \widetilde{u}_j \frac{\partial \widetilde{u}_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \widetilde{p}}{\partial x_i} + \nu \frac{\partial^2 \widetilde{u}_i}{\partial x_j \partial x_j} - \frac{\partial \tau_{ij}}{\partial x_j}, \quad \frac{\partial \widetilde{u}_i}{\partial x_i} = 0. \tag{2}$$

Most of the linear operations commute with spatially-homogeneous convolution while the nonlinear convection term $u_j \frac{\partial u_i}{\partial x_j}$ does not, resulting in an unclosed term τ, the so-called SGS stress given explicitly by

$$\tau_{ij} = \widetilde{u_i u_j} - \widetilde{u}_i \widetilde{u}_j. \tag{3}$$

The goal of the SGS modeling is to build a closure model to calculate the SGS stress from the filtered velocity field. For example, the well-known Smagorinsky model [1] provides $\tau = -2C_s\Delta^2|\widetilde{S}|\widetilde{S}$, where $\widetilde{S} = \frac{1}{2}(\nabla\widetilde{\mathbf{u}} + (\nabla\widetilde{\mathbf{u}})^T)$ is the rate of strain tensor, Δ is the grid scale, and C_s is the Smagorinsky constant.

The data-driven SGS model is commonly formalized as the optimization of the following regression problem:

$$\min_\theta \sum_n \left\| \phi_\theta(\widetilde{\mathbf{u}}^{(n)}) - \tau^{(n)} \right\|^2, \tag{4}$$

given the parametrized model ϕ_θ and the dataset $\{\widetilde{\mathbf{u}}^{(n)}, \tau^{(n)}\}$ where n denotes the sample. One can specify different choices of input features, and here we fix this to $\widetilde{\mathbf{u}}^{(n)}$ for simplification.

Strong evidence in previous work [2,3] and our experiments show that it is difficult to optimize this regression problem to a relative low error. In table 1. we illustrate the a-priori analysis using channel flow dataset for wall turbulence. The row denotes the input features where $\widetilde{\mathbf{u}}, \nabla\widetilde{\mathbf{u}}, \widetilde{S}, y^+$ are the filtered velocity, shear stresses, rate of strain, and wall unit. The outputs are the SGS stresses of dimension 9. We report the relative MSE on test dataset with models of different number of parameters.

Table 1: Summary of the a-priori error with different input features.

	# params	$\widetilde{\mathbf{u}}$	$\nabla\widetilde{\mathbf{u}}$	$\widetilde{S}$
NN, $[64]$	877	0.896	0.616	0.772
NN, $[64, 64]$	5037	0.884	0.601	0.775
NN, $[64, 64, 64]$	9708	0.923	0.586	0.774
XGBoost, $[100, 5]$	55910	0.888	0.618	0.759
XGBoost, $[1000, 5]$	558610	0.887	0.539	0.726
XGBoost, $[1000, 10]$	11055275	0.888	0.472	0.712

2.1 The issue of dataset imbalance

The first feature we observed from the data preprocessing is that the data pair of rate of strain and SGS stress is imbalanced. Namely, most of the data is concentrated in the region where both the rate of strain and the stresses are very small. To illustrate this point, we plot the histogram of one of the components of the shear stresses $\nabla\widetilde{\mathbf{u}}$ and stress tensor τ of NS equations and $\partial_{xx}u, \tau$ of KS equation in fig. 1. We observe that the data is highly concentrated around 0 for all cases, suggesting a possible multiscale structure of the dataset itself.

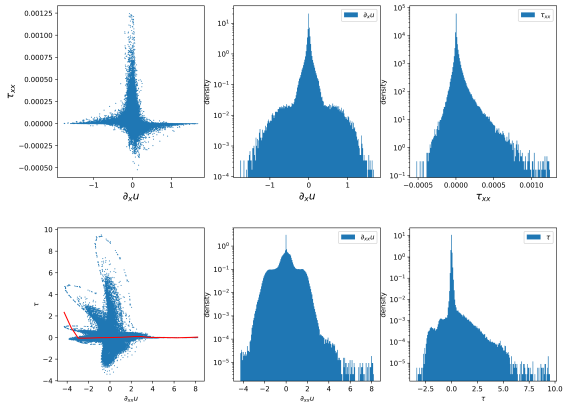

Fig. 1: We show both the scattering plot and the histogram of the SGS dataset for NS and KS equations. The upper (lower) row is the dataset for NS (KS) equations respectively. For NS equations, we only show the histogram of the first components of the shear $\nabla \mathbf{u}$ and SGS stresses τ, e.g. $\partial_x u, \tau_{xx}$. Notice that the y (density) axis is in log scale and there exists a peak around 0, indicating that a large potion of data has negligible magnitude.

2.2 The issue of multi-valuedness

Besides the data imbalance, we observe that the dataset is multivalue, which is more related to the cause of the large training error. The scattering plot of the data pair in both equations are shown in fig. 1, e.g. two figures on the left.

3. Generative modeling for the SGS modeling

To capture the multi-value nature of the SGS models, a direct solution is to model the SGS stresses as a probability distribution conditioned on the input features, instead of a single value in regression, i.e.

$$\tau = \phi_\theta(u) \quad \rightarrow \quad \tau \sim p_\theta(\cdot|u). \tag{5}$$

3.1 Conditional Gaussian model

We consider using the conditional Gaussian distribution for the SGS stresses, modeling the stresses as a Gaussian distribution depends on the input features. Instead of minimizing the mean squared error as in eq. (4), we use maximum likelihood principle to train the model, i.e. given a conditional model $p_\theta(\tau|u)$ we maximize the following function

$$\max_\theta \sum_{i=n}^{N} \log p_\theta(\tau^{(n)}|u^{(n)}). \tag{6}$$

Under the assumption that the parametrized family is Gaussian with learnable mean and variance, the loss function is simplified to:

$$\min_\theta \sum_{n=1}^{N} \frac{(\tau^{(n)} - \mu_\theta(u^{(n)}))^2}{2(\sigma_\theta(u^{(n)}))^2} + \log \sigma_\theta(u^{(n)}) \tag{7}$$

The comparison of this generative SGS model with the regression-based model is shown in table 2.

Table 2: Comparison of the a-priori and a-posteriori error of the method for SGS modeling of the KS equation. For a-priori error, we use the relative MSE for regression-based method while for generative modeling we use the expectation value of the log likelihood function.

	baseline	regression	gaussian, fix	gaussian sample
a-priori error	N/A	0.976	-2.173	-2.173
$\int (u-u_0)^2 dxdt$	1.524	2.036	1.720	1.597
$\overline{u} - \overline{u_0}$	9.901E-02	1.011E-01	3.870E-02	3.214E-02
$\overline{u^2} - \overline{u_0^2}$	-4.326E-01	-4.241E-01	-1.895E-01	-6.577E-02
a-priori error	N/A	0.987	-2.583	-2.583
$\int (u-u_0)^2 dxdt$	1.524	1.817	1.898	1.889
$\overline{u} - \overline{u_0}$	9.901E-02	8.012E-02	-1.242E-02	-1.543E-02
$\overline{u^2} - \overline{u_0^2}$	-4.326E-01	-3.466E-01	-1.018E-01	-1.296E-01

References

[1] Stephen B Pope. Turbulent flows. *Measurement Science and Technology*, 12(11):2020–2021, 2001.

[2] Chuwei Wang, Julius Berner, Zongyi Li, Di Zhou, Jiayun Wang, Jane Bae, and Anima Anandkumar. Beyond closure models: Learning chaotic-systems via physics-informed neural operators. *arXiv preprint arXiv:2408.05177*, 2024.

[3] Benjamin Sanderse, Panos Stinis, Romit Maulik, and Shady E Ahmed. Scientific machine learning for closure models in multiscale problems: A review. *arXiv preprint arXiv:2403.02913*, 2024.