14th International Symposium on Turbulence and Shear Flow Phenomena (TSFP14)
Heidelberg, Germany, July 28–31, 2026

FINDING SYMBOLIC SUBGRID SCALE CLOSURES BY COMBINING MACHINE LEARNING AND MULTISCALE FILTERING

Josef Hasslberger & Markus Klein
Department of Aerospace Engineering
University of the Bundeswehr Munich
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
josef.hasslberger@unibw.de

Maximilian Reissmann & Richard D. Sandberg
Department of Mechanical Engineering
University of Melbourne
Grattan Street, Parkville Victoria 3010, Australia

ABSTRACT

It has been recognized that both machine learning as well as explicit multiscale (or test) filtering are powerful techniques in the context of subgrid scale modeling for large eddy simulation. Hence, the main idea of this work is to combine the respective strengths of modern, i.e. data-driven, and classical modeling approaches. The potential of this combination is demonstrated by means of a-posteriori training based on a large number of coarse-grid Taylor–Green vortex simulations, where the subgrid scale stress is expressed in terms of basis tensors and invariants which are functions either of the resolved, high- or low-pass filtered velocity field. It is shown that the high-pass filtered velocity field yields model expressions with a superior performance on average. Furthermore, the data-driven models obtained are shown to generalize well for grids and Reynolds numbers unseen during training by using the high-pass filtered velocity field as an input, in contrast to the resolved or low-pass filtered velocity field.

MOTIVATION AND RESEARCH DESIGN

The application of machine learning techniques to turbulence modeling, e.g. in the context of computational fluid dynamics (CFD), has gained great momentum in recent years. In this context, gene expression programming (GEP) (Ferreira, 2001) emerged as an attractive technique which mimics biological evolution and yields human-readable symbolic expressions. In the framework of a CFD-driven training strategy, as sketched in Fig. 1, an efficient training configuration is essential due to the large number of individual CFD calculations to be conducted (of the order of 10^5 in the present application). Three-dimensional configurations are required in large eddy simulation (LES) due to the inherently three-dimensional nature of (resolved) turbulent fluctuations and the associated physical mechanisms such as vortex stretching, vortex breakdown and dissipation. As already demonstrated by Reissmann et al. (2021), the three-dimensional Taylor–Green vortex (TGV) ideally serves this purpose since it is mathematically well-defined and rich in physics at the same time (laminar-turbulent transition). Recently, it has been shown by Hasslberger et al. (2025) how an additional physical constraint, namely the asymptotic near-wall scaling of the shear stress, can be incorporated in this framework via semantic back-propagation – a popular machine learning technique pioneered by Rumelhart et al. (1986) and applied to genetic programming by Pawlak et al. (2015) and Reissmann et al. (2025).

Following this research line, the topic of the present work is to explore strategies on how to enrich the established CFD-GEP training framework by explicit multiscale (or test) filtering. The latter technique has demonstrated its great potential for subgrid scale modeling in the past. In this context and without any guarantee for completeness, it is worth mentioning Germano's seminal dynamic procedure to determine model constants (Germano et al., 1991), structural models based on the scale-similarity principle (Liu et al., 1994), the multiscale modeling framework (Hughes et al., 2001) or subgrid activity sensor enhanced eddy viscosity models (Hasslberger et al., 2021).

METHODOLOGY

Computational setup

The low-wavenumber initial condition of the Taylor–Green vortex (Brachet et al., 1983) reads

$$\begin{aligned} u(x,y,z) &= \cos(x)\sin(y)\sin(z), \\ v(x,y,z) &= -\sin(x)\cos(y)\sin(z), \\ w(x,y,z) &= 0, \end{aligned} \tag{1}$$

which can be well represented even on very coarse LES meshes. In the default configuration, the tri-periodic box of size $(2\pi)^3$ is discretized by 32^3 cells and the viscosity is adjusted such that the Reynolds number with respect to the initial state is Re=1600. The evolution of the flow in the considered non-dimensional time range, from $t = 0$ to $t = 25$, is indicated in Fig. 2. Starting from the quasi-laminar initial condition given by Eq. 1, the flow evolves towards fully developed turbulence via vortex stretching, breakdown and dissipation. During

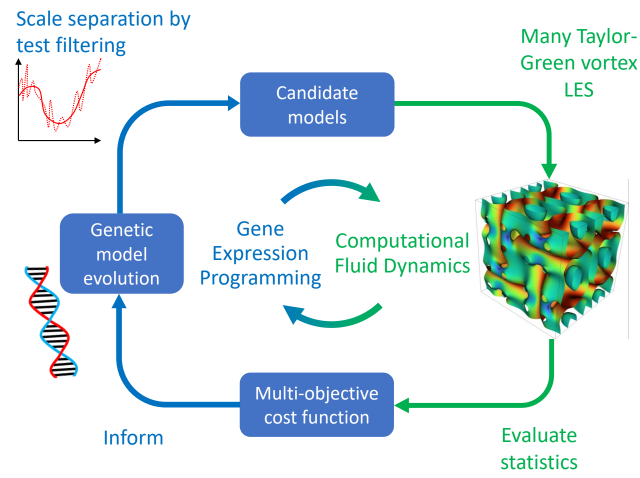

Figure 1. Schematic of the in-the-loop optimization scheme, coupling Gene Expression Programming (GEP) and Computational Fluid Dynamics (CFD). The evaluation of the multi-objective cost function involves Pareto ranking and the genetic model expressions are enriched by explicit multiscale (or test) filtering.

this laminar-turbulent transition process, ever smaller structures, down to the Kolmogorov scale, are produced, which necessitates SGS energy drain when insufficiently resolved. In order to guarantee optimal comparability, the highly resolved reference solution (DNS) and all coarse grid LES solutions have been generated with the same open-source code named PARIS (Aniszewski et al., 2021). It implements the unsteady incompressible Navier–Stokes equations based on a projection method. Spatial discretization follows a finite-volume scheme with second-order centered differences on a uniform staggered grid. Concurrently, time integration is realized by means of a second-order predictor-corrector technique.

Mathematical background

In the GEP framework used to find an explicit model, it is assumed that the subgrid scale stress tensor $\tau_{ij}^{sgs} = \overline{u_i u_j} - \overline{u}_i \overline{u}_j$ can be expressed as a function depending only on the LES filter width Δ as well as the grid-scale strain and rotation rate tensors,

$$\overline{S}_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right), \quad \overline{\Omega}_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} - \frac{\partial \overline{u}_j}{\partial x_i}\right). \tag{2}$$

For the convenient units-free handling in symbolic regression algorithms, $\overline{S}_{ij}$ and $\overline{\Omega}_{ij}$ are non-dimensionalized by an inverse time scale, $|\overline{S}|$, such that

$$s_{ij} = \frac{\overline{S}_{ij}}{|\overline{S}|}, \quad \omega_{ij} = \frac{\overline{\Omega}_{ij}}{|\overline{S}|}, \quad |\overline{S}| = \sqrt{\overline{S}_{mn}\overline{S}_{mn}}. \tag{3}$$

In the following, the overbar representing the filtering operation is omitted for the normalized quantities s_{ij} and ω_{ij} for conciseness. Any candidate model suggested by the GEP algorithm is formulated as

$$\tau_{ij}^{sgs} = -2\Delta^2 |\overline{S}|^2 \cdot f_{ij}\left(s_{ij}, \omega_{ij}\right) \tag{4}$$

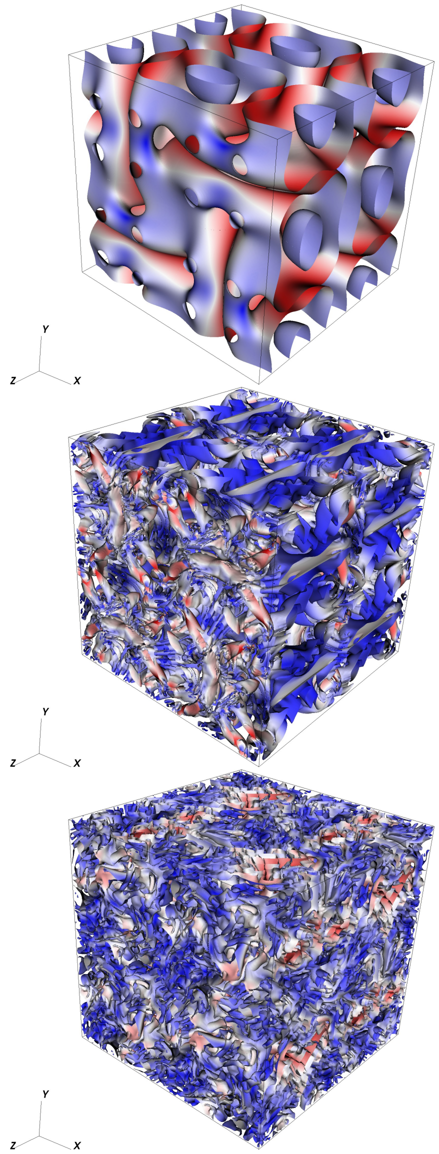

Figure 2. Taylor–Green vortex training configuration: Evolution of the reference solution for non-dimensional simulation times $t = 2.5, 7.5, 25$ (top to bottom); The iso-contours colored with the velocity magnitude are specified by $Q = 0$, where $Q = \frac{\partial u_j}{\partial x_i}\frac{\partial u_i}{\partial x_j}$ is the second invariant of the velocity gradient tensor.

which guarantees dimensional consistency. Making use of the Cayley–Hamilton theorem, the unknown function can be understood as a finite series of tensor basis functions t_{ij}^{α} with scalar coefficients G_α:

$$f_{ij} = \sum_{\alpha=1}^{n} G_\alpha \cdot t_{ij}^{\alpha}; \quad G_\alpha = G_\alpha(I_1, I_2, \ldots, I_m). \tag{5}$$

In order to make the optimization procedure simple, robust and computationally feasible, only the first four, up to quasi-quadratic, basis functions have been retained in the above series. The function space for f_{ij} is consequently spanned by the basis

$$
\begin{aligned}
t_{ij}^1 &= s_{ij}, \\
t_{ij}^2 &= s_{ik}\omega_{kj} - \omega_{ik}s_{kj}, \\
t_{ij}^3 &= s_{ik}s_{kj} - s_{mk}s_{km}\delta_{ij}/3, \\
t_{ij}^4 &= \omega_{ik}\omega_{kj} - \omega_{mk}\omega_{km}\delta_{ij}/3,
\end{aligned} \tag{6}
$$

and the scalar coefficients G_α are functions of the four invariants,

$$
\begin{aligned}
I_1 &= s_{mn}s_{nm}, \\
I_2 &= \omega_{mn}\omega_{nm}, \\
I_3 &= s_{km}s_{mn}s_{nk}, \\
I_4 &= \omega_{km}\omega_{mn}s_{nk},
\end{aligned} \tag{7}
$$

which are related to the physical mechanisms of dissipation of kinetic energy (I_1), enstrophy (I_2), generation of kinetic energy dissipation (I_3), and enstrophy production (I_4). The dimensionally consistent basis tensors T_{ij}^α can eventually be written as

$$
\begin{aligned}
T_{ij}^1 &= t_{ij}^1 \cdot |\overline{S}|, \\
T_{ij}^2 &= t_{ij}^2 \cdot |\overline{S}|^2, \\
T_{ij}^3 &= t_{ij}^3 \cdot |\overline{S}|^2, \\
T_{ij}^4 &= t_{ij}^4 \cdot |\overline{S}|^2.
\end{aligned} \tag{8}
$$

The suitability of candidate models can be assessed via the fitness of a function f_{ij} which is quantified by using the capability of multi-objective optimization (Waschkowski et al., 2022), based on the Non-dominated Sorting Genetic Algorithm (NSGA-II). Utilizing this approach, one objective conveys information about the discrepancy from the mean turbulent kinetic energy, and the other one the information from the dissipation rate evolution in the TGV training case. The selection is performed through Pareto domination.

Explicit Multiscale Filtering

Explicit low-pass filtering $\widehat{\phi}$ for any field quantity $\phi_{i,j,k}$ at the discrete location given by the index triple (i,j,k) is implemented according to Anderson & Domaradzki (2012):

$$
\widehat{\phi}_{i,j,k} = \sum_{l=-1}^{+1}\sum_{m=-1}^{+1}\sum_{n=-1}^{+1} b_l \cdot b_m \cdot b_n \cdot \phi_{i+l,j+m,k+n} \tag{9}
$$

This three-dimensional filter is the product of the convolution of three one-dimensional filters with coefficients $(b_{-1}, b_0, b_{+1}) = (c, 1-2c, c)$ where $c = 1/3$ is chosen here for sufficient scale separation. Hence, only the considered cell itself and direct neighbor cells ($l,m,n \in \{-1,0,+1\}$) are taken into account which makes the multiscale filtering procedure computationally inexpensive. Accordingly, high-pass filtering is realized via $\phi' = \phi - \widehat{\phi}$ and applied to the LES filtered quantities $\overline{\phi}$.

Apart from replacing the grid-scale velocity field $\overline{u}_i$ in Eq. 2 with either $\overline{u}_i'$ (high-pass filtering) or $\widehat{\overline{u}}_i$ (low-pass filtering), the mathematical framework remains identical in the present work. Beyond that, it is conceivable to implement a different treatment for basis tensors and invariants, and this will also be discussed at the conference (not shown here for brevity).

MODEL TRAINING AND ASSESSMENT

Figure 3 provides an overview of the evolutionary training process spanning 50 generations. It is obvious from Fig. 3 that the minimum cost function is consistently smaller (i.e. higher model fitness) when using the high-pass filtered velocity field as the model input in comparison to the low-pass filtered version and the one using only the grid filter. In this way, the subgrid scale model is sensitized towards small-scale resolved velocity fluctuations which is beneficial for the accurate prediction of laminar-turbulent transition as present in the initial stage of the Taylor–Green vortex benchmark. In other words, high-pass filtering may help to realize a model expression which is robust against Reynolds number variation. Conversely, low-pass filtering leads to consistently lower model fitness. Furthermore, when using the high-pass filtered input, a distinct slope in the evolutionary trajectory is visible up to generation 50. Hence, in this case, there is a high likelihood of finding even better models when resuming the training beyond the 50 generations currently considered.

Among all runs of the evolutionary training, the best models for the high-pass filtered, the low-pass filtered and the default unfiltered velocity field input read:

$$
\begin{aligned}
\tau_{ij}^{High-pass}(\overline{u}_i') = -2\Delta^2[&(I_3 + (I_3 - 0.1)(I_3 - 0.02)) \\
&(2I_1 + 0.9I_2 + I_3 + 1.02) \cdot T_{ij}^3 - (I_1 - 0.5) \cdot T_{ij}^4]
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
\tau_{ij}^{Low-pass}(\widehat{\overline{u}}_i) = -2\Delta^2[&-(I_1 - I_4)(I_3 + 0.5) \cdot T_{ij}^2 \\
&+ (I_3^2 - I_4) \cdot T_{ij}^3 + (0.6I_1 - 0.06) \cdot T_{ij}^4]
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
\tau_{ij}^{Default}(\overline{u}_i) = -2\Delta^2[&-(0.5I_3 + 0.19) \cdot T_{ij}^2 \\
&+ 0.1I_1I_3(I_1 + 1)(2I_1 + 2I_2 - 1) \cdot T_{ij}^3]
\end{aligned} \tag{12}
$$

It is worth noting that basis tensor T_{ij}^1 remains unselected in this training configuration.

The observed higher or lower model fitness is also evident from the temporal evolution of the domain-averaged turbulent kinetic energy and its dissipation in the Taylor–Green vortex as shown in Fig. 4. The peak dissipation, at around 8 non-dimensional simulation times, is considerably better reproduced by the GEP model with the high-pass input compared to the low-pass input, and this is directly reflected in the training process via the multi-objective cost function. However, the improvement through high-pass filtering comes at the cost of minor unphysical oscillations of the dissipation curve – yet these are not jeopardizing the stability of the simulation. In general, the data-driven GEP models show superior agreement with the highly resolved reference results compared to the no-model LES as well as well-established eddy viscosity models from the literature (Smagorinsky and Sigma model).

Data-driven methods usually reproduce well the reference data under the given training conditions (here, Reynolds number of Re=1600 on a 32^3 grid). However, a truly predictive model must generalize well beyond the training conditions. Accordingly, the best GEP models are tested for two additional Reynolds numbers (Re=400 and Re=3000) and grids (64^3 and 128^3) unseen in the training. Figure 5 shows the same TGV

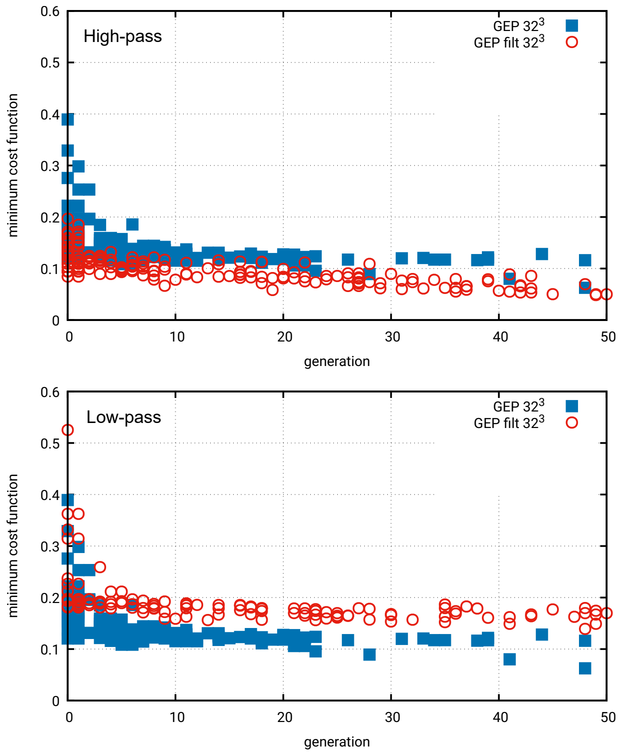

Figure 3. Evolution of the minimum cost function for GEP model training using high-pass filtered input data (red markers in the top subfigure) and low-pass filtered input data (red markers in the bottom subfigure). Blue markers (in both subfigures) indicate the corresponding evolution without explicit multiscale filtering for comparison.

statistics as before and demonstrates a clearly superior behavior of the high-pass model when the Reynolds number is varied. In particular, the peak dissipation scales correctly with the Reynolds number, i.e. higher peak dissipation for higher Reynolds number, whereas the opposite is the case for the low-pass model. Likewise for the grid variation depicted in Fig. 6, only the high-pass model shows the sought-after behavior. For all resolutions tested, the high-pass GEP model agrees well with the highly resolved reference result.

CONCLUDING REMARKS

In summary, the present study gives an impression of the benefits and limitations of combining explicit multiscale filtering with machine learning of symbolic subgrid scale closures. Based on Gene Expression Programming, the adopted CFD-driven a-posteriori training strategy directly includes the effect of numerical discretization errors and also evaluates the stability of the candidate model expressions. This is especially relevant in LES due to the strong interaction between numerical and modeling errors.

In the challenging Taylor-Green vortex test case, featuring laminar-turbulent transition, consistently higher subgrid model fitness has been achieved by using a high-pass filtered velocity field as the model input in comparison to a low-pass filtered or unfiltered velocity field. Furthermore, it can be concluded that the high-pass filtered velocity input yields clearly superior generalization properties of the data-driven models for Reynolds numbers and grids unseen in the training.

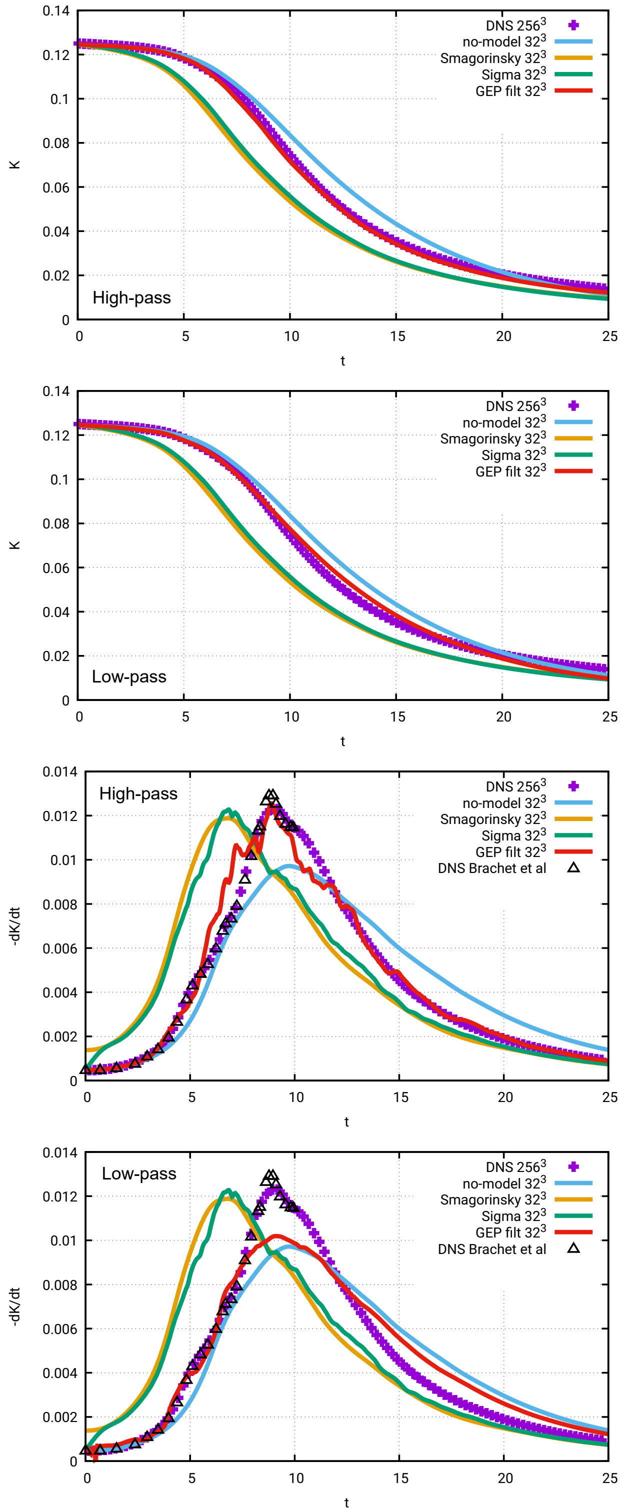

Figure 4. Temporal evolution of the domain-averaged turbulent kinetic energy K and its dissipation $-dK/dt$ in the Taylor–Green vortex for the best GEP models (red lines) using high-pass filtered input data and low-pass filtered input data. Results from a highly resolved reference solution as well as standard models from the literature (Smagorinsky and Sigma model) are also included for comparison. The numbers in the legend indicate the grid resolution.

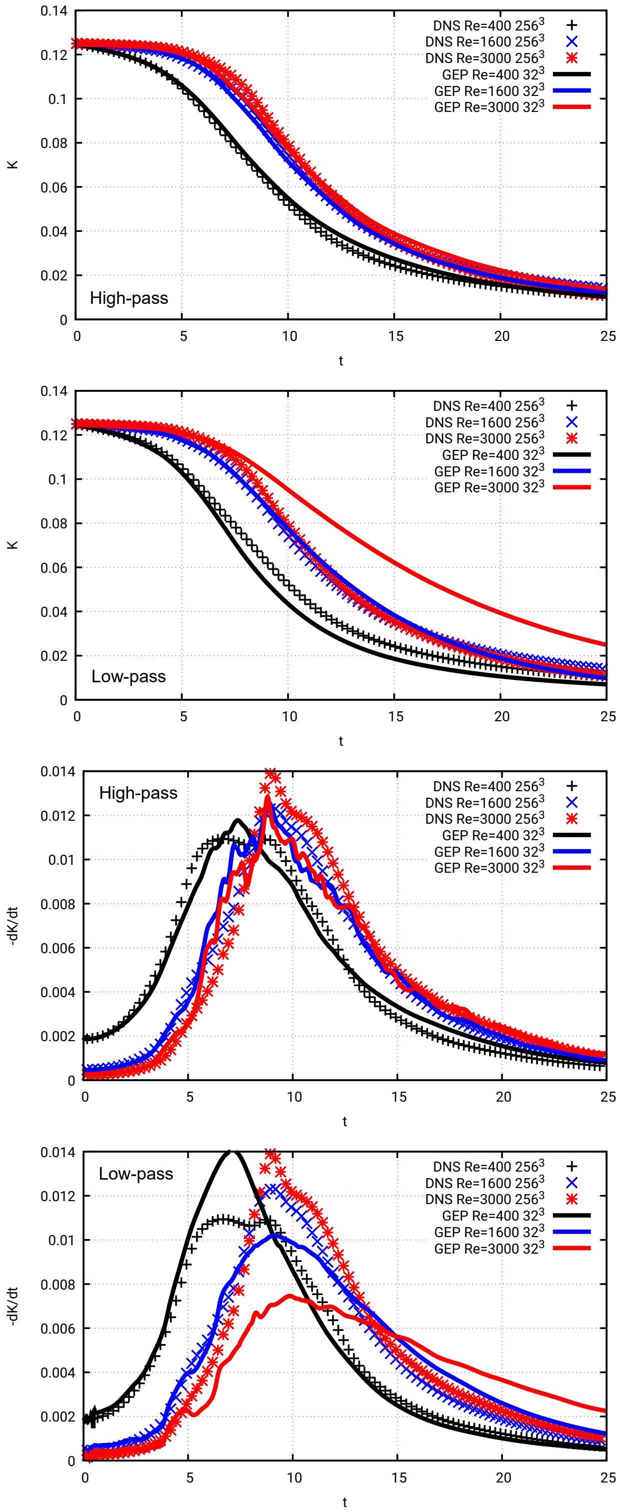

Figure 5. Temporal evolution of the domain-averaged turbulent kinetic energy K and its dissipation $-dK/dt$ in the Taylor–Green vortex for the best GEP models using high-pass filtered input data and low-pass filtered input data. The a-posteriori training has been performed for a Reynolds number of Re=1600 (blue) whereas Re=400 (black) and Re=3000 (red) represent conditions unseen in the training. Marker points indicate the corresponding highly resolved reference solutions.

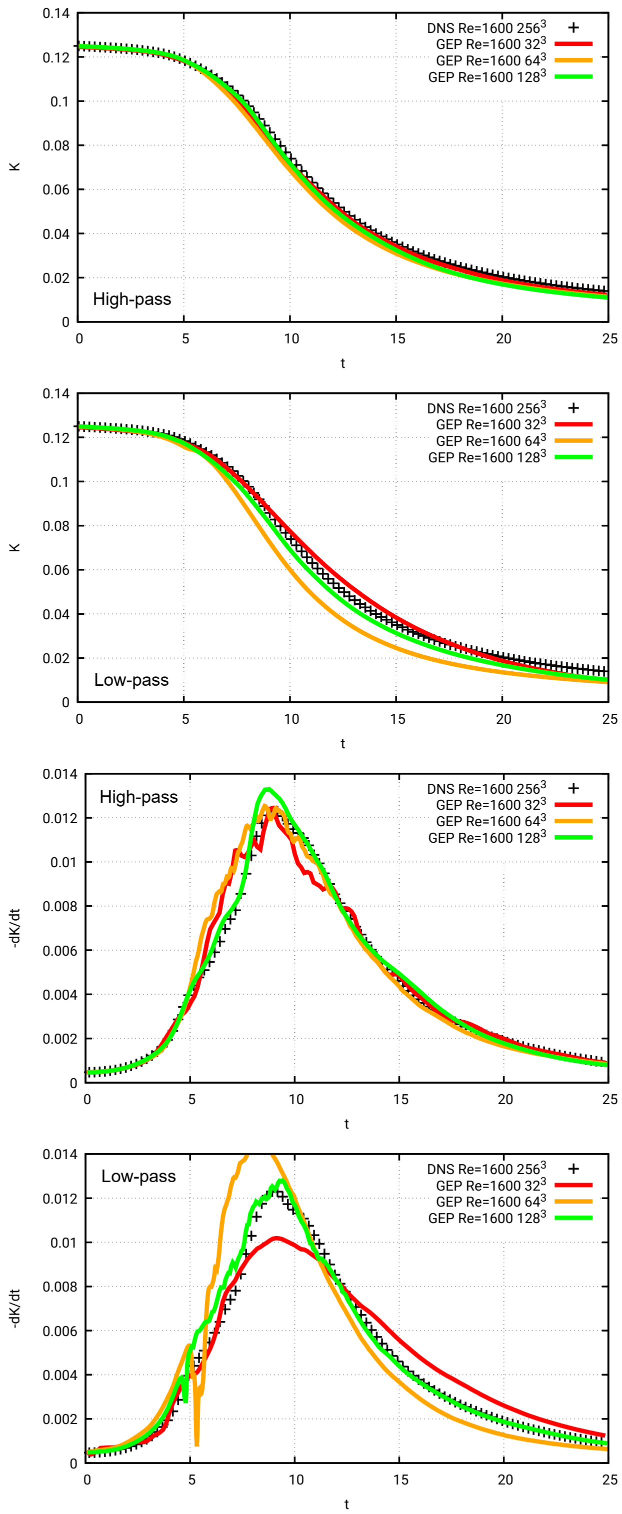

Figure 6. Temporal evolution of the domain-averaged turbulent kinetic energy K and its dissipation $-dK/dt$ in the Taylor–Green vortex for the best GEP models using high-pass filtered input data and low-pass filtered input data. The a-posteriori training has been performed on the 32^3 grid (red) whereas grids of 64^3 (orange) and 128^3 (green) represent conditions unseen in the training. Black markers indicate the highly resolved reference solution.

REFERENCES

Anderson, B. W., and Domaradzki, J. A., 2012, ”A subgrid-scale model for large-eddy simulation based on the physics of interscale energy transfer in turbulence”, Physics of Fluids, 24(6).

Aniszewski, W., Arrufat, T., Crialesi-Esposito, M., Dabiri, S., Fuster, D., Ling, Y., and Zaleski, S., 2021, ”Parallel, robust, interface simulator (PARIS)”, Computer physics communications, 263, 107849.

Brachet, M.E., Meiron, D.I., Orszag, S.A., Nickel, B.G., Morf, R.H., and Frisch, U., 1983, ”Small-scale structure of the Taylor–Green vortex”, Journal of Fluid Mechanics, 130, 411-452.

Ferreira, C., 2001, ”Gene expression programming: a new adaptive algorithm for solving problems”, Complex Systems, 13(2), 87–129.

Germano, M., Piomelli, U., Moin, P., and Cabot, W. H., 1991, ”A dynamic subgrid-scale eddy viscosity model”, Physics of fluids A: Fluid dynamics, 3(7), 1760-1765.

Hasslberger, J., Engelmann, L., Kempf, A., and Klein, M., 2021, ”Robust dynamic adaptation of the Smagorinsky model based on a sub-grid activity sensor”, Physics of Fluids, 33(1).

Hasslberger, J., Reissmann, M., Sandberg, R. D., and Klein, M., 2025, ”Exploiting semantic backpropagation to impose near-wall scaling constraints in machine-learned symbolic subgrid-scale closures”, Physics of Fluids, 37(3).

Hughes, T. J., Mazzei, L., Oberai, A. A., and Wray, A. A., 2001, ”The multiscale formulation of large eddy simulation: Decay of homogeneous isotropic turbulence”, Physics of Fluids, 13(2), 505-512.

Liu, S., Meneveau, C., and Katz, J., 1994, ”On the properties of similarity subgrid-scale models as deduced from measurements in a turbulent jet”, Journal of Fluid Mechanics, 275, 83-119.

Nicoud, F., Toda, H. B., Cabrit, O., Bose, S., and Lee, J., 2011, ”Using singular values to build a subgrid-scale model for large eddy simulations”, Physics of Fluids, 23(8).

Pawlak, T.P., Wieloch, B., Krawiec, K., 2015, ”Semantic backpropagation for designing search operators in genetic programming”, IEEE Transactions on Evolutionary Computation 19, 326–340.

Reissmann, M., Hasslberger, J., Sandberg, R.D., and Klein, M., 2021, ”Application of gene expression programming to a-posteriori LES modeling of a Taylor Green vortex”, Journal of Computational Physics, 424, 109859.

Reissmann, M., Fang, Y., Ooi, A. S., and Sandberg, R.D., 2025, ”Constraining genetic symbolic regression via semantic backpropagation”, Genetic Programming and Evolvable Machines, 26(1), 12.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986, ”Learning representations by back-propagating errors”, Nature, 323(6088), 533–536.

Smagorinsky, J., 1963, ”General circulation experiments with the primitive equations: I. The basic experiment”, Monthly weather review, 91(3), 99-164.

Waschkowski, F., Zhao, Y., Sandberg, R.D., and Klewicki, J., 2022, ”Multi-objective CFD-driven development of coupled turbulence closure models”, Journal of Computational Physics, 452, 110922.