Supplementary File for "Mapping Network-coordinated Stacked Gated Recurrent Units for Turbulence Prediction"

Zhiming Zhang, Shangce Gao, *Senior Member, IEEE*, MengChu Zhou, *Fellow, IEEE*, Mengtao Yan, and Shuyang Cao

THIS part gives the supplementary materials to support the research in the main body.

A. Hyper-Parameters Analysis

To comprehensively investigate the network structure and learning strategy in MSU, a sensitivity analysis of the hyper-parameters is performed. Specifically, we consider the loss function coefficient in $\mathcal{L}_c$, the structure of a prediction network, and that of a mapping network, which are summarized in Table S.I. For the loss function's coefficient β, according to the analysis of MSU's learning process, $\beta \geq 0.5$ is necessary for mapping the network. Therefore, the three fixed values are selected for comparison, ranging from 0.9 to 0.5. Furthermore, a dynamic linearly varying β is also taken into account. It changes from 0.1 to 0.9 depending on the progress of the training, formulated as:

TABLE S.I
THE COMPARISON OF DIFFERENT HYPER-PARAMETERS

Hyper-parameter	Value
Loss function coefficient	dynamic, 0.95, 0.9, 0.7, 0.5
Size of prediction network hidden layers	[512], [512, 256], [512, 512, 256], [512, 1024, 256], [512, 1024, 512], [512, 512, 256, 256]
The first convolutional kernel size of mapping network	32, 64,128

$$\beta = 0.1 + 0.8 \times \frac{E}{\hat{E}} \tag{1}$$

where E and $\hat{E}$ are current training epoch count and the maximum epoch count, respectively. For the structure of the prediction network, not only stacked but single GRU cells are included in the comparison, containing a total of three different network architectures. For the structure of a mapping network, the size of the first convolutional kernel determines the whole network. In total, we consider three different structures. We recombine the three terms into 90 different hyper-parameter combinations ($5 \times 6 \times 3 = 90$) to perform a gird search [1], [2] and each combination runs 10 times to find the best hyper-parameter combination of MSU.

Fig. S.1. shows the box-whisker plots of mapping mask activation degree M, R^2 (spatial-averaged), and mean (temporal-averaged). From the plot of M, we can find that if $\beta \leq 0.7$ the method cannot perform the prediction based on just one velocity point, but necessarily needs several velocity points. Besides, we rank the other combinations among all metrics (R^2, RMSE, MAE, Mean, and standard deviation (Std)) and mark the top-ranked one as red in the figure. The best hyper-parameters for MSU are: $\beta = 0.9$, the prediction network structure is [512, 1024, 512], and the first convolutional kernel size of the mapping network is 64. This set of hyper-parameters not only has outstanding best and median values in 10 tests but also has good stability. It is worth emphasizing that this parameter setting consistently provides satisfactory prediction results based on only one velocity point.

REFERENCES

[1] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[2] P. Zhang, S. Shu, and M. C. Zhou, "An online fault detection method based on SVM-grid for cloud computing systems," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 2, pp. 445–456, 2018.

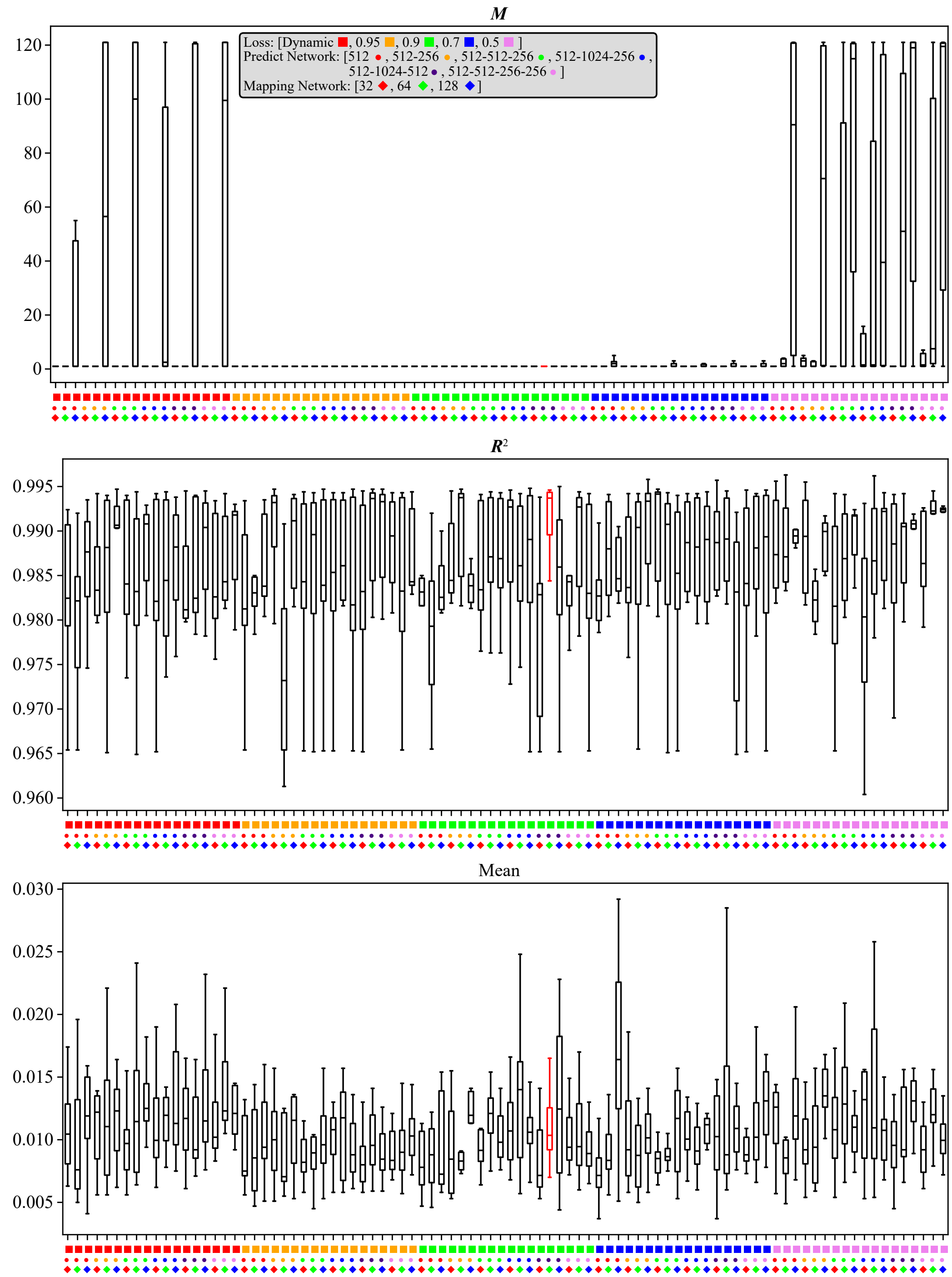

Fig. S1. The box-whisker plot of three performance metrics in the hyper-parameter sensitivity analysis for MSU.