Supplementary Materials
Deep ocean learning of small scale turbulence

Ali Mashayek[1], Nick Reynard[1], Fangming Zhai[1], Kaushik Srinivasan[2], Adam Jelley[3], Alberto Naveira Garabato[4], Colm-cille P. Caulfield[5]

[1]Imperial College London, UK
[2]University of Califorina Los Angeles, USA
[3]University of Edinburgh, UK
[4]University of Southampton, UK
[5]University of Cambridge, UK

S1. Global Observational Surveys Figs S1a-d show the coverage of the global observational surveys that provide T, S, Z data (in addition to other fields) that can be used to infer estimates of turbulent mixing either through finescale parameterizations (Polzin et al., 2014) or through data-driven methods as in this study. These field programs do not contain direct turbulent measurements. Of relevance to this work is the hydrographic surveying component of these experiments, which provide high-quality conductivity-temperature-pressure profiles to construct a climatological temperature-salinity-depth database. Figs S1e,f show the high-resolution topography data that are key to turbulence prediction, due to the importance of the bottom boundary in generating propagating waves as well as non-propagating boundary turbulence. Gravity data provide coarser topographic information than the direct echo-sounding surveys, which cover only 30% of the seafloor, but are extending their coverage at an accelerating rate. High-resolution seafloor mapping is also commonly provided by deep-ocean surveying research cruises, and is integrated in global topographic data (e.g. https://www.gebco.net/https://www.gebco.net/).

S2. Neural Network Architecture and Training Standard FNNs consist of a series of 'layers' of neurons that are hierarchically modified by matrix multiplication and vector addition of learned parameters and acted upon by a simple nonlinear function. Thus if $(h_0, h_1, h_2...h_L)$ represent the NN layers, with $h_0 = x$ being the input and $h_L = y$ the output, then the NN can be written by the recurrence relation for the ℓ^{th} layer as

$$h_\ell = f(W_{\ell-1}\, h_{\ell-1} + b_{\ell-1}), \tag{1}$$

where $W_{\ell-1}$ and $b_{\ell-1}$ are the learnable weight matrix and the bias vector acting on the $(\ell-1)^{th}$ hidden layer and $f(\cdot)$ is a simple nonlinear function here chosen to be the Swish activation function (Ramachandran et al., 2017) [$f(x) = x\sigma(x)$ where $\sigma(x)$ is the Sigmoid function] that is chosen over the standard ReLU activation function owing to its smoothness and in our case, improved predictive accuracy.

Residual networks have a simple architectural modification in that the layer-wise recurrence relation now takes the form

$$h_\ell = h_{\ell-1} + f(W_{\ell-1}\, h_{\ell-1} + b_{\ell-1}), \tag{2}$$

so that each neural layer is an add-on onto the previous hidden layer. Resnets have been shown to have smoother gradient flow during backpropagation allowing for deeper layers. More importantly, however, Resnets have been demonstrated to be implicitly composed of ensembles of shallower neural networks (Veit et al., 2016) which can result in substantially improved expressivity and accuracy compared to standard NNs. The specific choice of the Resnet used for the results in this manuscript has 7 layers with 120 neurons in each layer for a total of around 100,000 parameters in the NN. Each hidden layer is also subject to dropout regularization to prevent overfitting (with a layerwise dropout probability of 0.2) though the primary regularization in our approach is implicit and due to learning-rate annealing (see below).

Training is done through the AdamW optimizer (Adam with weight decay) with a weight decay parameter of 10^{-4}. A cyclical cosine learning rate annealing (Loshchilov &

Hutter, 2016) is employed with annealing cycle, $T_{cycle} = 5$ epochs. In other words the learning rate changes every 5 epochs starting from its largest value of 0.0035, decreasing towards 0 as a cosine function, and jumping back suddenly to 0.0035 every sixth epoch. Training for each run is performed for about 3000 epochs. This rapid decrease of the learning rate followed by sudden increase (also called a 'warm restart') leads to faster learning and provides a strong regularization. We use a mean-square error loss function but record the best value of R^2 metric on the test data and the corresponding model parameters during training (because lowest MSE loss does not always correspond to the best R^2 value). The regularization offered by the cyclical rate learning rate annealing with short T_{cycle} is sufficiently strong so that overfitting is not observed even as we train for larger epochs (up to 10,000) with larger NNs (up to 12 layers). This training approach is extremely robust even in small data regimes and needs minimal hyperparameter search.

References

Davis, R. E., Talley, L. D., Roemmich, D., Owens, W. B., Rudnick, D. L., Toole, J., ... Barth, J. A. (2019). 100 years of progress in ocean observing systems. *Meteorological Monographs*, *59*, 3–1.

GEOTRACERS. (2019). Geotracers. *https://www.geotraces.org/*. Retrieved from `https://www.geotraces.org/`

GO-SHIP. (2018). Go-ship. *http://www.go-ship.org/*. Retrieved from `http://www.go-ship.org/`

Gouretski, V., & Koltermann, K. P. (2004). {WOCE} global hydrographic climatology. *Berichte des BSH*, *35*, 1–52.

Loshchilov, I., & Hutter, F. (2016). SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Polzin, K. L., Garabato, A. C. N., Huussen, T. N., Sloyan, B. M., & Waterman, S. (2014). Finescale parameterizations of turbulent dissipation. *Journal of Geophysical Research: Oceans*, *119*(2), 1383–1419.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Sandwell, D. T., Müller, R. D., Smith, W. H., Garcia, E., & Francis, R. (2014). New global marine gravity model from cryosat-2 and jason-1 reveals buried tectonic structure. *Science*, *346*(6205), 65–67.

Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, *29*.

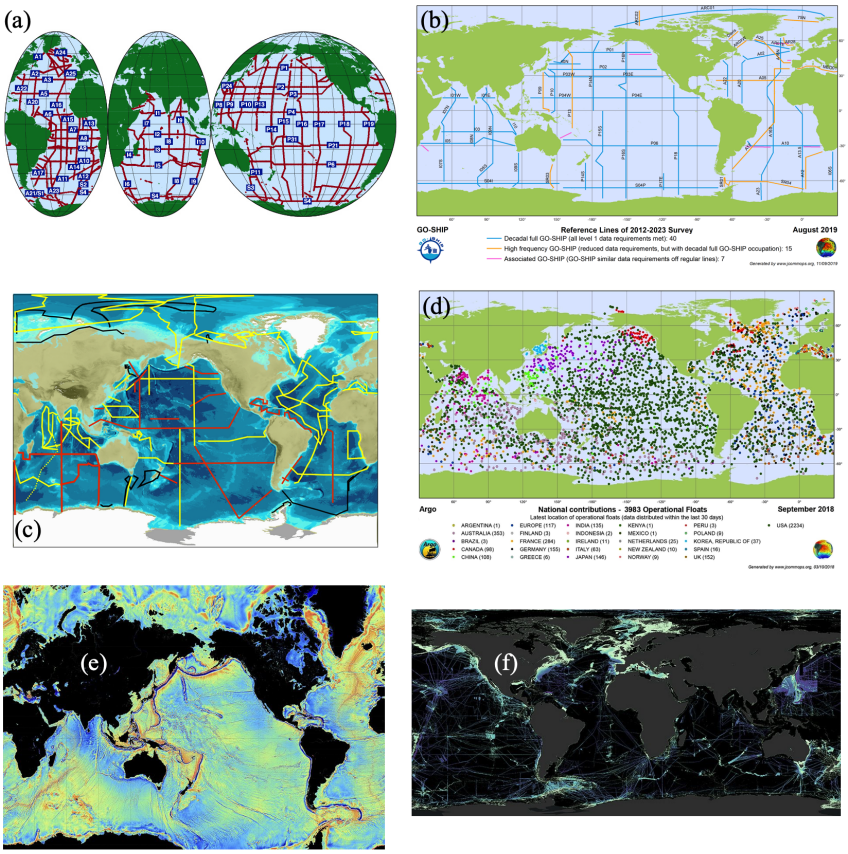

Figure 1. International surveys have provided invaluable hydrographic and bathymetric information required to quantify oceanic turbulent processes. (a) WOCE Hydrographic Program survey stations [1985–97](Gouretski & Koltermann, 2004; Davis et al., 2019). (b) GO-SHIP hydrographic sections [GO-SHIP 2018](GO-SHIP, 2018; Davis et al., 2019). (c) GEOTRACES sections [from 2018](GEOTRACERS, 2019; Davis et al., 2019). (d) Global Argo array coverage [as of 2018](Davis et al., 2019).(e) Satellite-measured marine gravity, revealing the ocean bathymetric features.(Sandwell et al., 2014) (f) Black regions represent the areas yet to be measured with echo-sounders, whereas lines represent already sampled regions ($\sim$ 20% as of 2020) [from NIPPON FOUNDATION-GEBCO SEABED 2030 PROJECT].

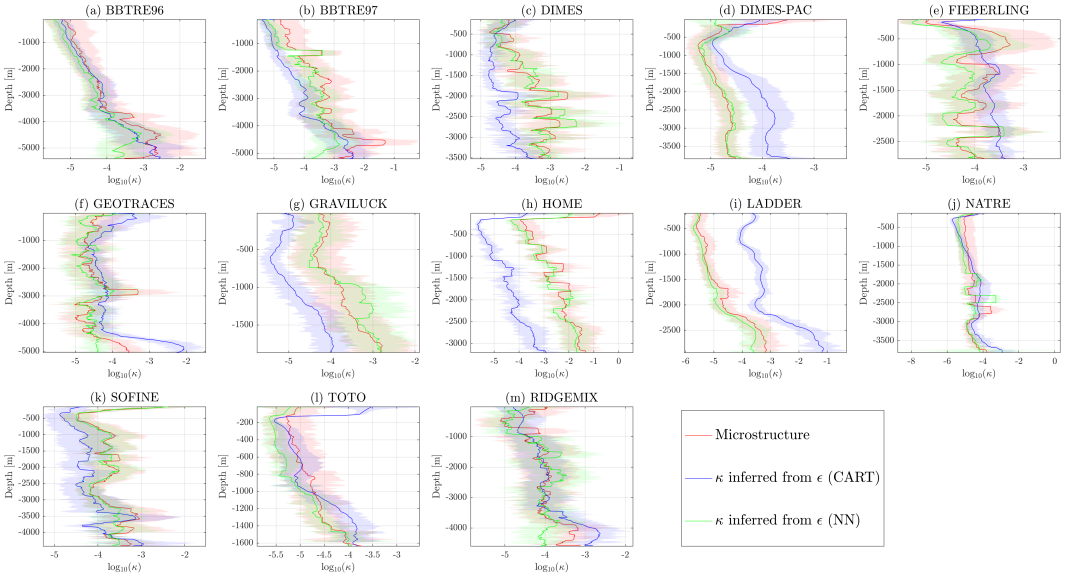

Figure 2. Same as Fig. 4 in the main text, but here the models are trained to predict ϵ and then κ is inferred from that prediction using Eq. (1) with $\Gamma = 0.2$.