Deep ocean learning of wave-induced turbulence

Ali Mashayek[1,1], Fangming Zhai[2,2], Nick Reynard[2,2], Adam Jelley[3,3], Colm Caulfield[4,4], Alberto C. Naveira Garabato[5,5], and Kashik Srinivasan[6]

[1]Imperial College London
[2]Imperial College
[3]University of Edinburgh
[4]University of Cambridge
[5]University of Southampton
[6]University of California Los Angeles

November 30, 2022

Abstract

Turbulent mixing at centimetre scales is an essential component of the ocean's meridional overturning circulation and its associated global redistribution of heat, carbon, nutrients, pollutants and other tracers. Whereas direct turbulence observations in the ocean interior are limited to a modest collection of field programs, basic information such as temperature, salinity and depth (T,S,Z) is available globally. Here, we show that supervised (deep) machine learning algorithms, informed by physical understanding, can be trained on the existing turbulence data to develop skillful predictions of the key properties of turbulence from T,S,Z and topographic data. This constitutes a promising first step toward a hybrid physics - artificial intelligence approach to parameterize turbulent mixing in climate-scale ocean models.

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

Key Points:

- Machine learning can be used to infer ocean turbulent mixing from basic seawater and geometric properties.
- The machine learning models are trained based on limited available direct turbulence measurements.
- The trained models can be applied to data from global observational programs, which do not sample turbulence directly.

Corresponding author: Ali Mashayek, mashayek@ic.ac.uk

Abstract

Turbulent mixing at the sub-meter scale is an essential component of the ocean's meridional overturning circulation and its associated global redistribution of heat, carbon, nutrients, pollutants and other tracers. Whereas direct turbulence observations in the ocean interior are limited to a modest collection of field programs, basic information such as temperature, salinity and depth is available globally. Here, we show that supervised machine learning algorithms can be trained on the existing turbulence data to develop skillful predictions of the key properties of turbulence from T, S, Z and topographic data. This constitutes a promising first step toward a hybrid physics-artificial intelligence approach to parameterization of turbulent mixing in climate models.

Plain Language Summary

Ocean turbulence plays an important role in sustaining the general ocean circulation and in the mixing of heat, carbon, nutrients, and other processes within the ocean interior. Turbulent mixing is technically challenging to measure and is often inferred from measurable quantities using parameterizations that are based on numerous simplifying assumptions about the physics of turbulence. In this study, we show that artificial intelligence (more specifically, various machine learning algorithms) can be successfully employed to infer turbulent mixing from quantities measured routinely by global observational programs.

Introduction

Turbulent mixing across density surfaces (i.e. diapycnal mixing) in the ocean interior is key to sustaining the meridional overturning circulation and its global regulation of heat, carbon and nutrient distributions, as well as other climatically and environmentally important tracers (Talley et al., 2016). Such turbulence is primarily excited at the ocean surface by winds, or at the bottom boundary via flow impingement on topography (Garabato & Meredith, 2022). The spatio-temporal variability of turbulence makes its measurement especially challenging. However, turbulence can leave an imprint on vertical temperature (T) and salinity (S) profiles obtained from hydrographic surveys. T, S and depth (Z) are regularly sampled through global international programs, such as ship-based efforts like WOCE (Gouretski & Koltermann, 2004), GO-SHIP (GO-SHIP, 2018), GEOTRACES(GEOTRACERS, 2019) or globally-distributed floats deployed by the Argo Program (Argo, 2000) (see Supplementary Materials for a visual summary, and Davis *et al.* (2019) for a review (Davis et al., 2019)). While turbulence characteristics may be inferred from these T, S, Z data (Polzin et al., 2014; Whalen et al., 2012), such estimates involve many assumptions and uncertainties.

The gold standard in measuring turbulence in the ocean interior is represented by ship-deployed microstructure profiler observations, which include concurrent sampling of T, S and Z, but are limited in number due to their technical complexity and cost (Shroyer et al., 2018). In this study, we train machine learning models on a unique collection of observations from microstructure field programs enabling prediction of turbulence characteristics based on T, S, Z and topographic data, rendering our approach applicable to major global surveys that do not measure turbulence directly. Our aim is to demonstrate that such predictions from microstructure-trained physics-inspired machine-learning models yield better estimates for dynamically-significant quantities than classical finestructure parameterizations.

Physics of Turbulence

A key property of density-stratified ocean turbulence is the significantly enhanced rate (as compared to molecular diffusion) at which it mixes density and tracers in the vertical (Thorpe, 2005). In observations and climate models, such mixing is often encapsulated

in a turbulent diffusion coefficient (or diffusivity for short) defined as

$$\kappa = \Gamma \frac{\varepsilon}{N^2}, \tag{1}$$

where Γ is a coefficient that determines the fraction of the energy available to turbulence that contributes to mixing (Peltier & Caulfield, 2003), ε is the rate of dissipation of turbulent kinetic energy (due to viscosity of seawater)[1], and $N = \sqrt{-[g/\rho_0]\partial\rho/\partial z}$ is the buoyancy frequency (Osborn, 1980). While Γ is known to be variable (Mashayek & Peltier, 2013; Mashayek et al., 2017; Gregg et al., 2018), for the purpose of this study it suffices to consider it a constant, specifically 0.2, in line with operational physical oceanography (Gregg et al., 2018; Mashayek et al., 2021). N can be directly inferred from measurements of T, S and Z through the construction of the vertical density gradient, a characteristic density ρ_0, and the gravitational acceleration g. On the other hand, the dissipation rate ε, as it is determined from the strain-rate tensor, cannot be inferred from T, S, Z (which are available from global observational programs) and is best inferred from microstructure profilers, that measure spatial gradients of velocity. In this study, we show that machine learning models can be trained on microstructure data to predict ε (or directly predict κ) based on T, S, Z, and height above the bottom (Hab). [2] This allows for global inference of κ from observational surveys, thereby providing a route for direct application to climate models that assimilate data from such surveys (e.g. Forget et al. (2015); Verdy and Mazloff (2017)).

The Training Dataset

We employ a global dataset of microstructure profiles compiled by the Climate Process Team on internal wave–driven ocean mixing (MacKinnon et al., 2017). Fig. 1 shows the location of the field measurements, spanning a wide range of geographic locations, depths, and turbulence-inducing physical processes. A sample microstructure transect from the DIMES experiment is shown in panel *c* (more specifically, transect T1 in Fig. 5a). Fig. 1 also provides the list of the field experiments, and the fraction of the total data associated with each experiment. The data are available at `https://microstructure.ucsd.edu/`, and data description and relevant references may be found in Waterhouse et al. (2014). The same dataset was employed by Cael and Mashayek (2021) to show that the data 'collapses' on a seemingly universal log-skew-normal statistical distribution. This finding motivated the present study by suggesting that such universality might be detectable through data-driven methods. Together, the experiments in Fig. 1 contain over 700 full-depth microstructure profiles, binned into 10 m vertical bins (amounting to $\sim 2\times10^5$ data points). The concurrent measurements of ε, T, S, Z in this dataset allow for the construction of the aforementioned predictor list (i.e. the list of features used in training) used to predict ε and κ. More specifically, neutral density is calculated from T, S, Z, latitude, and longitude information (Jackett & McDougall, 1997), the local depth for each profile is looked up from the global bathymetric map of Sandwell et al. (2014), and height above is then calculated by subtracting the sample depth from the local depth.

[1] More precisely, $\varepsilon = \nu \frac{\partial u_i'}{\partial x_j} \frac{\partial u_i'}{\partial x_j}$, where u_i' represent perturbation velocity components (i.e. departures from the mean flow), x_i represents the three Cartesian dimensions, and the overbar represents an 'appropriate' averaging.

[2] We found that inclusion of both Z and Hab is crucial as they represent the distance from the top and bottom boundaries, both of which are turbulence generation sites. Knowledge of Hab requires topographic data, which has become increasingly more accurate in recent decades thanks to advanced satellite-based gravity measurements and deep-ocean echo-sounding records (Sandwell et al., 2014) (see Supplementary Materials Fig. S1).

Machine Learning Models

Fig. 2 illustrates the overall flowchart for the research presented here: assembling the training datasets (as shown in Fig. 1); training two machine learning models with distinctly different underlying algorithms; assessing the models' skills (as shown in Fig. 3); and independent verification of the models through their application to individual field programs (as shown in Figs 4 and 5). This section describes the construction of the two machine learning algorithms.

Classification And Regression Trees (CART)

We employ CART, one of the most common machine learning predictive models (Wu et al., 2008; Breiman et al., 1984). The method uses a decision tree to connect observations of a parameter of interest (represented in the branches) to predictions about its value (represented in the leaves). When applied to target variables that take continuous values (such as ε or κ in this study), such decision trees are referred to as regression trees. Additionally, we employ an ensemble method, bootstrap aggregating, to improve the stability and accuracy of the decision tree algorithm, reduce variance, and avoid overfitting. Bootstrap aggregated decision trees (hereafter bagging trees) construct multiple trees by repeatedly re-sampling the training data with replacements, and voting the trees for a consensus prediction (Breiman, 1996).

Figs 3a,b show the application of the bagging tree to the training microstructure dataset (from which $\log_{10}(\varepsilon)$, our prediction target, is calculated). The model was trained based on 10 cross-validation k-folds of all data across 13 field experiments. This method involves splitting the dataset into equally sized 'k' number of groups, or 'folds', and taking it in turn to use each group as the test data while the rest of the data is used to train the model, with an average of the results being adopted. A k-fold validation approach is useful when input data is limited, and ensures that every data point is used within the training and test dataset, hence reducing bias when compared to other methods. The fits in Figs 3a,b are satisfactory, with a coefficient of determination (R^2) of 0.83 for $\log_{10}(\varepsilon)$ and 0.84 for $\log_{10}(\kappa)$. [3] To analyze further the quality of the agreement between predictions and data, panels $e-h$ display the cumulative contribution of various predictors to increases in R^2 and decreases in the mean squared error (MSE).

We consider two sets of predictors, with nearly equal skills. First we consider T, S, their gradients, Z, Hab, and latitude. Secondly, we just use $\log_{10}(N^2)$, Z, Hab, and latitude.[4] While the two sets show similar skills, it is worth noting that the former contains more raw information about the temperature and salinity structures (which get combined into one parameter once N is calculated). It is conceivable that in regions of the ocean where salinity structures play a key role in turbulence generating processes (e.g. double diffusion in the Arctic Ocean; Middleton et al. (2021)), retaining T, S and their derivatives may prove fruitful. We postpone the investigation of application of our methodology to such regions to future work.

It is worth noting that we also tried another standard choice, namely the Least Squares Boost (LSBoost) algorithm, as an alternative ensemble learning method. LSBoost is a gradient boosting method in which the mean squared error is chosen as the cost function (Breiman et al., 1984). While we found LSBoost to outperform bagging tree for a smaller number of features (up to 3), bagging tree was superior for the number of features employed herein,

[3] R^2 is a statistical metric of how well the regression predictions approximate the real data, and so is a measure of the goodness of fit of a model.

[4] Since quantities like ε, κ, and N vary over orders of magnitude, employing their logarithms renders the training algorithms more efficient.

and thus is our method of choice. Finally, we note that application of a linear regression model to the dataset proved entirely futile.

Neural Networks

As an entirely different approach, we also train neural networks with the same data. Specifically, we use a fully-connected feed-forward neural network (FNN), also referred to as a Multi-layer Perceptron (MLP) in the broader ML community. Standard FNNs consist of an input layer, an output layer and multiple hidden layers in between (Goodfellow et al., 2016) but we use a slight modification of this FNN architecture by making the hidden layers actually *residual* layers. Unlike standard hidden layers, which recursively perform operations on the previous layer, residual layers are added on to the main input-to-output information flow. (Such additions are also referred to as *skip connections* (He et al., 2016).) NNs with predominantly residual layers, or Resnets, have been found to outperform direct NNs, not just on the class of problems relevant to this study (Gorishniy et al., 2021), but across almost all modalities in AI/ML in general (Drozdzal et al., 2016; Vaswani et al., 2017) and hence residual layers are correspondingly ubiquitous features of most modern NNs. Each hidden layer combines the (learned) features of the previous layer to build up a non-linear transformation of the input predictors to predict turbulence properties (ε and κ) in the output layer. Adding additional layers to make the network deeper incorporates more parameters to be learned, which allows for a more flexible mapping between the easily measurable predictors and the less widely available turbulence properties. Typically, adding more parameters requires more data to learn an effective generalizable mapping without overfitting. However, we use a specific training algorithm called stochastic gradient descent with warm restarts (Loshchilov & Hutter, 2016) that provides a strong implicit regularization, essentially eliminating the issue of overfitting, even in low data regimes. A 10-fold cross-validation algorithm is used to ensure coverage of the entire dataset using the trained NN models. The Supplementary Materials contain more information on the resnet-FNN model architecture as well as details of the training and optimization procedure.

Figs 3c,d show that the deep neural network is also skillful in predicting both ε and κ. Deep learning algorithms like neural networks require less human intervention compared to more traditional machine learning algorithms (e.g. the bagging tree), and so generally have larger data requirements and their performance increases more strongly with the size of data. This makes the high R^2 values for NNs in Fig 3b particularly promising, given the limited nature of the training data compared to data sizes typically employed in deep learning. Thus, investment in extending the training data through a community effort appears worthwhile. We note that while CART seems to give a slightly higher R^2 than NN in Fig. 3, as we will show next, NN proves more skillful in capturing the vertical patterns of turbulence for individual experiments when they are considered separately.

Application to Individual Datasets

Fig. 4 shows the results of separate analyses for each of the 13 different field programs listed in Fig 1. Importantly, the data from each experiment are excluded from training of the models before the models are applied to it. While both NN and CART show skills in predicting the patterns and, in some cases, the order of magnitude adequately, NN is clearly superior in both respects. While Fig. 4 shows predictions based on models trained to infer κ directly, we have also repeated the exercise based on models trained to predict ε, and then inferred κ from that prediction using Eq. 1 with $\Gamma = 0.2$. The outcome, shown in Supplementary Materials Fig. S2, is qualitatively similar, although, importantly, the direct prediction of κ is more skilled at predicting the turbulence-induced diffusivity in the vicinity of seafloor. This superiority of 'direct' estimation of diffusivity is significant since such turbulent mixing is key to the upwelling of the deep waters formed and sunk at high

latitudes, a process necessary for closure of the oceanic meridional overturning circulation (de Lavergne et al., 2022).

Indirect inferences of turbulent mixing from T and S finestructure, practically the only alternative when microstructure data is unavailable, can be inaccurate by as much as two orders of magnitude (Polzin et al., 2014). Furthermore, such parameterizations are based on somewhat restrictive assumptions regarding the nature of the underlying turbulence-generating processes. Thus, the accuracy of NN showcased in Fig. 4, in light of its agnosticism towards the underlying physics, is appealing. To further highlight this point, in Fig. 5 we assess the skill of CART and NN for the data sampled along three transects (shown in Fig. 5a) as a part of the DIMES experiment. For these transects, both direct (from microstructure profilers) and indirect finestructure-based parameterizations of ε and κ were reported in Sheen et al. (2013), allowing for testing our models against conventional finestructure parameterizations (Figs 5b-g). Both NN and CART outperform the finestructure parameterization, particularly for κ (which is ultimately the parameter of interest). It is worth noting that the study of Sheen et al. (2013) is one of the more successful applications of finestructure parameterization; examples of much larger disagreements between finestructure and microstructure estimates abound in the literature.

Discussion & Outlook

The primary message of this study is that AI can indeed be successfully employed to use data from global observational programs, which lack direct turbulence measurements, to predict small scale turbulent mixing in the ocean, and in particular, more accurately than conventional finestructure parameterizations. More specifically, this study implies that the knowledge of parameters most basic to turbulence, i.e. finescale density stratification, distance from turbulence-generating boundaries, and latitude, suffice to leading order to obtain an estimate of the turbulence intensity and the associated turbulent (density) diffusivity.

There are numerous factors that can contribute to the misfits between the predictions and the data. Three important ones are: (I) the percentage of the training and validation data can vary significantly between the experiments (as shown in Fig. 1); (II) the relevance of the underlying physics in each experiment to the rest of the data used for training might be limited; (III) ocean mixing is not entirely 'local' in nature, e.g. waves generated thousands of kilometers away can contribute to mixing, and no such information was included in our training by construction (de Lavergne et al., 2019). Factor (I) can only be addressed through application of AI to larger datasets. In particular, the success of deep learning directly scales with the data size, and what was achieved in this study lies at the lower bound of the data volume required. Our analyses show that while the bagging tree algorithm converges to the optimal performance once a few hundreds of profiles are considered, the NN algorithm does not show such convergence and retains a large standard deviation even when all profiles are included. Thus, further community efforts are required to pull turbulence datasets together and subject them to the consistent high levels of quality control and grid interpolations. Furthermore, adding microstructure sensors to global observational endeavors (such as the Argo float program), while ambitious, is within reach and conceivable in the coming decades (Roemmich et al., 2019). Factor (II) will naturally advance as our physical understanding of ocean turbulence keeps progressing. A conscious effort towards connecting such physical understanding to data-driven parameterizations is required. Addressing factor (III) is more readily achievable in the near future, as it will require inclusion of theoretical estimates of local and non-local energy injected to internal waves from various sources (winds, tides, etc.) in training algorithms. In summary, we have demonstrated here that AI provides a valuable tool to harness our observational, theoretical and statistical knowledge of ocean turbulence to direct the development of a next-generation 'smart' turbulence parameterization for climate models.

Acknowledgments

The authors report no conflict of interests. The authors thank Kathy Sheen for providing the data from Sheen et al. (2013) for the purpose of constructing Fig. 5, and Lois Baker and two anonymous reviewers for constructive comments.

Data Availability Statement

The microstructure data employed for the training of the Machine Learning algorithms may be obtained from `https://microstructure.ucsd.edu/` by locating the names of the experiments in Figure 3; also see (Waterhouse et al., 2014) and (MacKinnon et al., 2017) for further information. The AI algorithms will be shared via an online depository upon acceptance of the manuscript.

References

Argo, G. (2000). Argo float data and metadata from global data assembly centre (argo gdac). *SEANOE*.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, *37*(15), 237–251.

Cael, B., & Mashayek, A. (2021). Log-skew-normality of ocean turbulence. *Physical Review Letters*, *126*(22), 224502.

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.

de Lavergne, C., Falahat, S., Madec, G., Roquet, F., Nycander, J., & Vic, C. (2019, 5). Toward global maps of internal tide energy sinks. *Ocean Modelling*, *137*, 52–75. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S1463500318302890` doi: 10.1016/j.ocemod.2019.03.010

de Lavergne, C., Groeskamp, S., Zika, J., & Johnson, H. L. (2022). The role of mixing in the large-scale ocean circulation. *Ocean Mixing*, 35–63.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., & Pal, C. (2016). The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications* (pp. 179–187). Springer.

Forget, G., Campin, J.-M., Heimbach, P., Hill, C. N., Ponte, R. M., & Wunsch, C. (2015). ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation.

Garabato, A. N., & Meredith, M. (2022). Ocean mixing: oceanography at a watershed. In *Ocean mixing* (pp. 1–4). Elsevier.

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (`http://www.deeplearningbook.org`)

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, *34*.

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.

Gregg, M., D'Asaro, E., Riley, J., & Kunze, E. (2018, 1). Mixing Efficiency in the Ocean. *Annual Review of Marine Science*, *10*(1), 443–473. Retrieved from `http://www.annualreviews.org/doi/10.1146/annurev-marine-121916-063643` doi: 10.1146/annurev-marine-121916-063643

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp.

770–778).

Jackett, D. R., & McDougall, T. J. (1997, 2). A Neutral Density Variable for the World's Oceans. *Journal of Physical Oceanography*, *27*(2), 237–263. doi: 10.1175/1520-0485(1997)027⟨0237:ANDVFT⟩2.0.CO;2

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

MacKinnon, J. A., Zhao, Z., Whalen, C. B., Waterhouse, A. F., Trossman, D. S., Sun, O. M., ... others (2017). Climate process team on internal wave–driven ocean mixing. *Bulletin of the American Meteorological Society*, *98*(11), 2429–2454.

Mashayek, A., Caulfield, C. P., & Alford, M. H. (2021). Goldilocks mixing in oceanic shear-induced turbulent overturns. *Journal of Fluid Mechanics*, *928*, A1.

Mashayek, A., & Peltier, W. R. (2013). Shear-induced mixing in geophysical flows: does the route to turbulence matter to its efficiency? *Journal of Fluid Mechanics*, *725*, 216–261.

Mashayek, A., Salehipour, H., Bouffard, D., Caulfield, C. P., Ferrari, R., Nikurashin, M., ... Smyth, W. D. (2017). Efficiency of turbulent mixing in the abyssal ocean circulation. *Geophysical Research Letters*, *44*(12), 6296–6306.

Middleton, L., Fine, E., MacKinnon, J., Alford, M., & Taylor, J. (2021). Estimating dissipation rates associated with double diffusion. *Geophysical Research Letters*, *48*(15), e2021GL092779.

Osborn, T. R. (1980). Estimates of the local rate of vertical diffusion from dissipation measurements. *Journal of Physical Oceanography*, *10*, 83–89.

Peltier, W. R., & Caulfield, C. P. (2003). Mixing efficiency in stratified shear flows. *Annual Rev. Fluid Mech.*, *35*, 135–167.

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.

Roemmich, D., Alford, M., Claustre, H., Johnson, K., King, B., Moum, J., et al. (2019). On the future of argo: an enhanced global array of physical and biogeochemical sensing floats. front. *Mar. Sci*, *6*, 439.

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.

Sheen, K. L., Brearley, J. A., Naveira Garabato, A. C., Smeed, D. A., Waterman, S., Ledwell, J. R., ... others (2013). Rates and mechanisms of turbulent dissipation and mixing in the Southern Ocean: Results from the Diapycnal and Isopycnal Mixing Experiment in the Southern Ocean (DIMES). *Journal of Geophysical Research: Oceans*, *118*(6), 2774–2792.

Shroyer, E. L., Nash, J. D., Waterhouse, A. F., & Moum, J. N. (2018). Measuring Ocean Turbulence. In (pp. 99–122). Springer, Cham. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-319-66493-4_6` doi: 10.1007/978-3-319-66493-4{_}6

Talley, L. D., Feely, R. A., Sloyan, B. M., Wanninkhof, R., Baringer, M. O., Bullister, J. L., ... Zhang, J. Z. (2016, 1). Changes in Ocean Heat, Carbon Content, and Ventilation: A Review of the First Decade of GO-SHIP Global Repeat Hydrography. *Annual Review of Marine Science*, *8*, 185–215. Retrieved from `www.go-ship.org` doi: 10.1146/annurev-marine-052915-100829

Thorpe, S. A. (2005). The turbulent ocean. *Cambridge University Press*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Verdy, A., & Mazloff, M. R. (2017). A data assimilating model for estimating Southern Ocean biogeochemistry. Retrieved from `http://hycom.org` doi: 10.1002/2016JC012650

Waterhouse, A. F., MacKinnon, J. A., Nash, J. D., Alford, M. H., Kunze, E., Simmons,

H. L., ... others (2014). Global patterns of diapycnal mixing from measurements of the turbulent dissipation rate. *Journal of Physical Oceanography*, *44*(7), 1854–1872.

Whalen, C., Talley, L., & MacKinnon, J. (2012). Spatial and temporal variability of global ocean mixing inferred from argo profiles. *Geophysical Research Letters*, *39*(18).

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... others (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1–37.

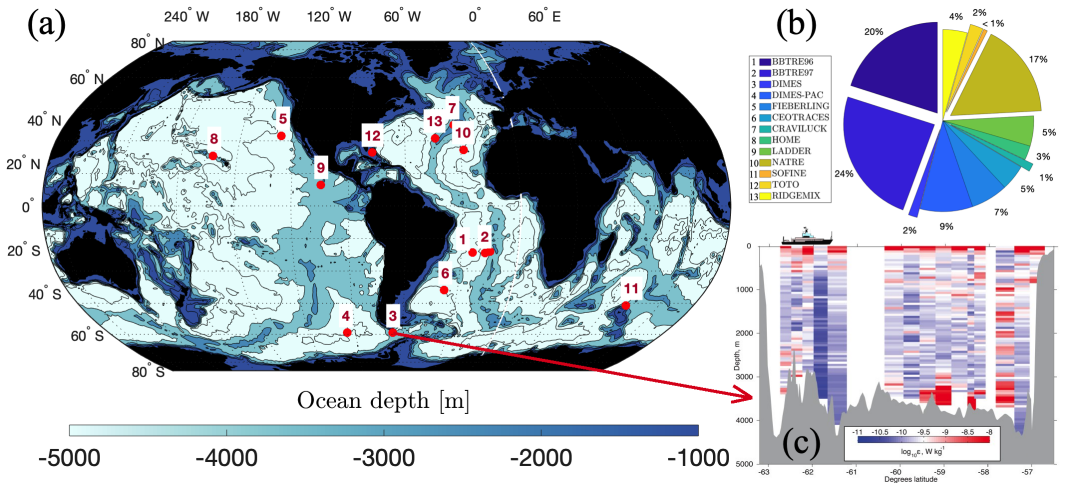

Figure 1. Direct turbulence measurements can be used to train machine learning algorithms to predict turbulent mixing where direct measurements are not available. (a) Location of the field programs that include direct measurements of turbulence (specifically, turbulent kinetic energy dissipation rate ε from microstructure profilers) along with co-located temperature, salinity and depth sampling. (b) The experiments' name and associated contributions to the total data. More details about the data sources are available at `https://microstructure.ucsd.edu/`(Waterhouse et al., 2014) and in the Supplementary Materials. The data contains a total of ∼700 profiles, with ε binned into 10 m vertical bins. (c) A sample transect of microstructure data from the DIMES experiment (transect T1 in Fig. 5); from Sheen et al. (2013).

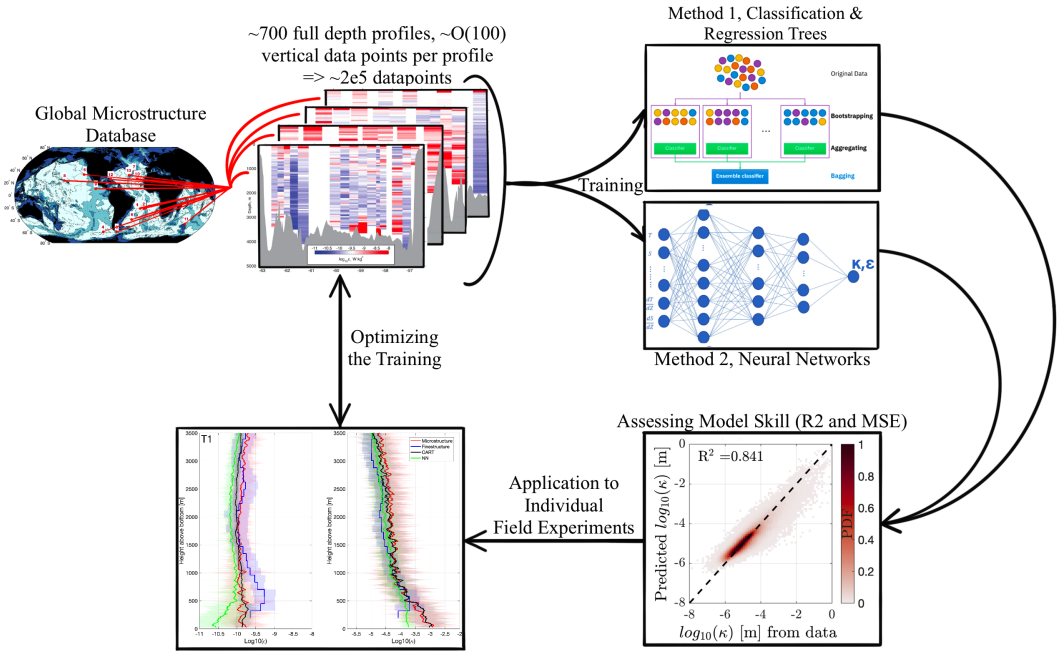

Figure 2. Two distinctly different machine learning algorithms can successfully reproduce turbulent mixing estimates in agreement with microstructure data.
A flowchart, illustrating the sequence of data assembly, training, model skill assessment, application to original data, verification, and fine-tuning. The source for the sample microstructure profiles shown in the top row is Sheen et al. (2013)– see Fig. 5 for details. The source for the CART diagram in the top row is https://en.wikipedia.org/wiki/Bootstrap_aggregating. Note that CART and NN are not applied sequentially, but are independent algorithms.

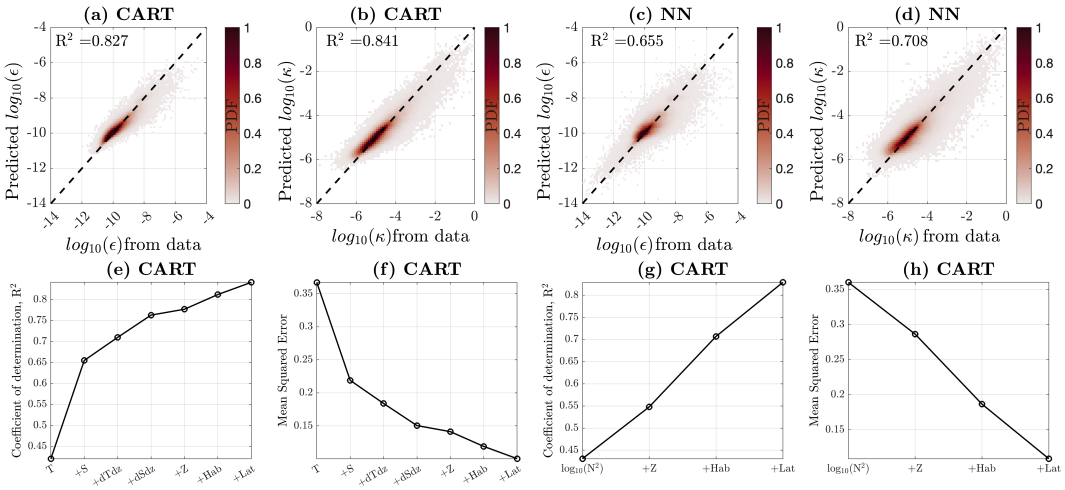

Figure 3. Machine learning can successfully fit the global microstructure data based on few predicting features.

(a-d) Bivariate histograms (in the form of a probability density function, PDF) of predicted rate of dissipation of turbulent kinetic energy (ε; [m^2/s^3]) and turbulent diffusivity (κ; [m^2/s]) based on use of the Classification And Regression Tree algorithm (CART; panels *a*,*b*), and Neural Networks (NN; panels *c*,*d*), versus the actual data. Both CART and NN models are validated using k-fold validation with 10 folds to avoid overfitting (see main text). (e-f) Cumulative contributions from each of the training features to the increase in the coefficient of determination (R^2) and the decrease in the mean squared error (MSE). (g-h) same as panels e, f but for a smaller number of features. All the datasets shown in Fig. 1 are employed in this figure.

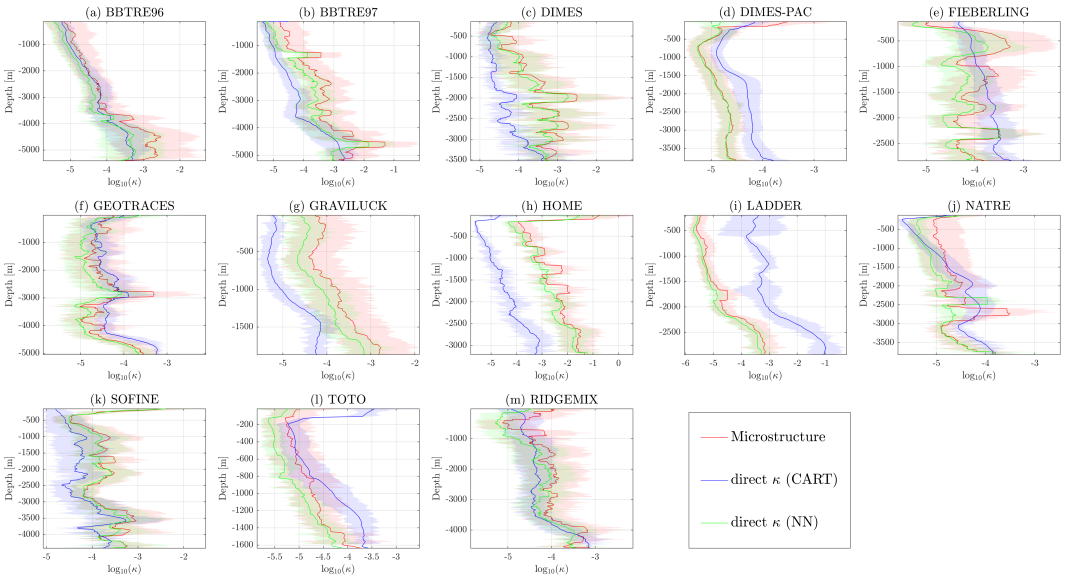

Figure 4. Predictions for individual field programs also show promise. Comparison of the predictions of the machine learning algorithms (CART and NN) to each of the 13 field programs introduced in Fig 1. For each case, the solid lines represent the mean over all the profiles in that experiment and the corresponding shadings represent standard deviation. Note that individual predictions are made for each profile of each experiment, before averaging. For each experiment, the models were trained based on the data from all other 12 experiments, **excluding** the data from the given experiment itself, to avoid overfitting. This figure shows results from models trained to predict the turbulent diffusivity κ directly. A similar plot, showing qualitatively the same level of success, is included in the Supplementary Materials in which the models predict ϵ and κ is constructed using (1). Direct inference of κ seems to be better for predicting turbulence near the seafloor.

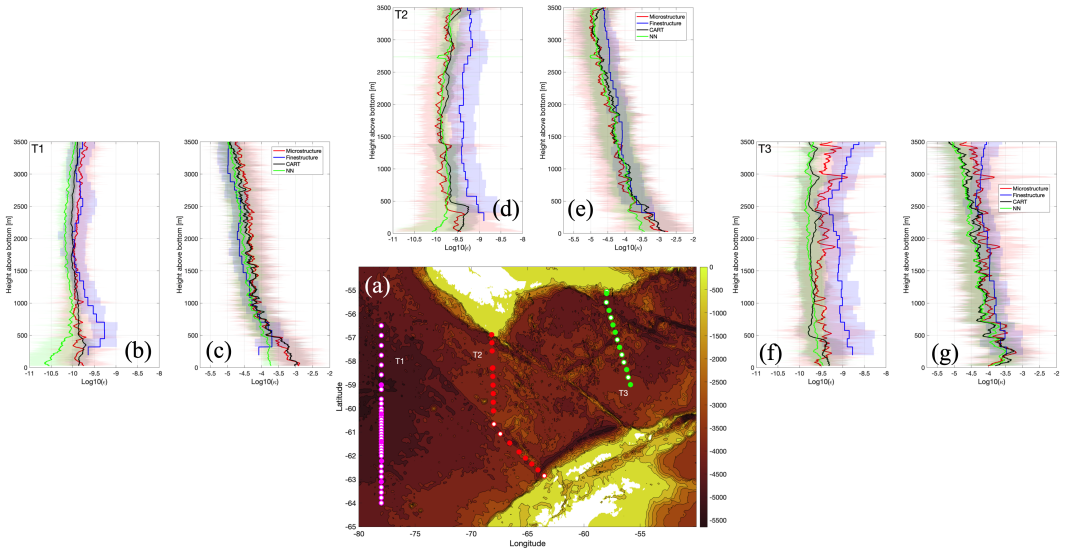

Figure 5. Machine learning competes well against physics-based parameterizations. Comparison of the predictions from machine learning against finestructure parameterization for three transects of the DIMES experiment. (a) The Drake Passage of the Southern Ocean, with the three cruise transects marked. Each circle represents a sampling. Transect T1 is the most western. plotted in magenta. Filled circles mark locations where microstructure data was taken, along with T, S, Z (from which finestructure based estimates of are inferred). The circles with a white filling do not include microstructure sampling. The means over all profiles for each transect are calculated for microstructure-based, finestructure-based, and machine learning-based (both CART and NN) estimates of ε and κ for transect T1 in (b,c), transect T2 in (d,e), and transect T3 in (f,g). The plots are in height-above-bottom (Hab) coordinate, due to strong bottom-enhanced topographically-induced turbulence in the Drake Passage. Microstructure and finestructure estimates are from Sheen et al. (2013).

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*.

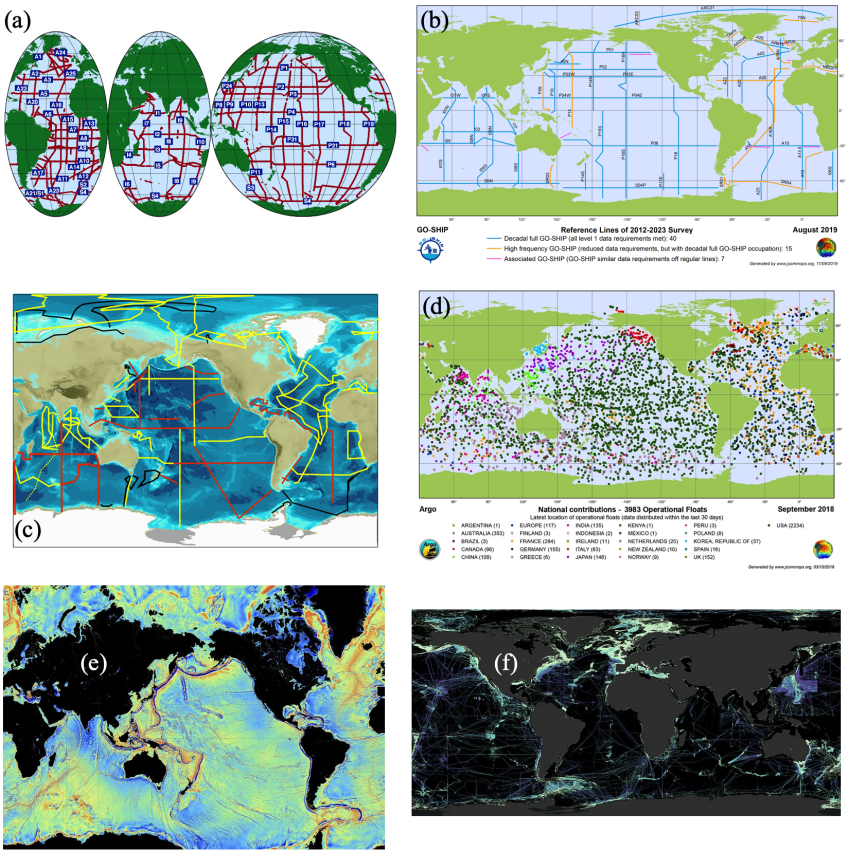

Figure 1. International surveys have provided invaluable hydrographic and bathymetric information required to quantify oceanic turbulent processes. (a) WOCE Hydrographic Program survey stations [1985–97](Gouretski & Koltermann, 2004; Davis et al., 2019). (b) GO-SHIP hydrographic sections [GO-SHIP 2018](GO-SHIP, 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].

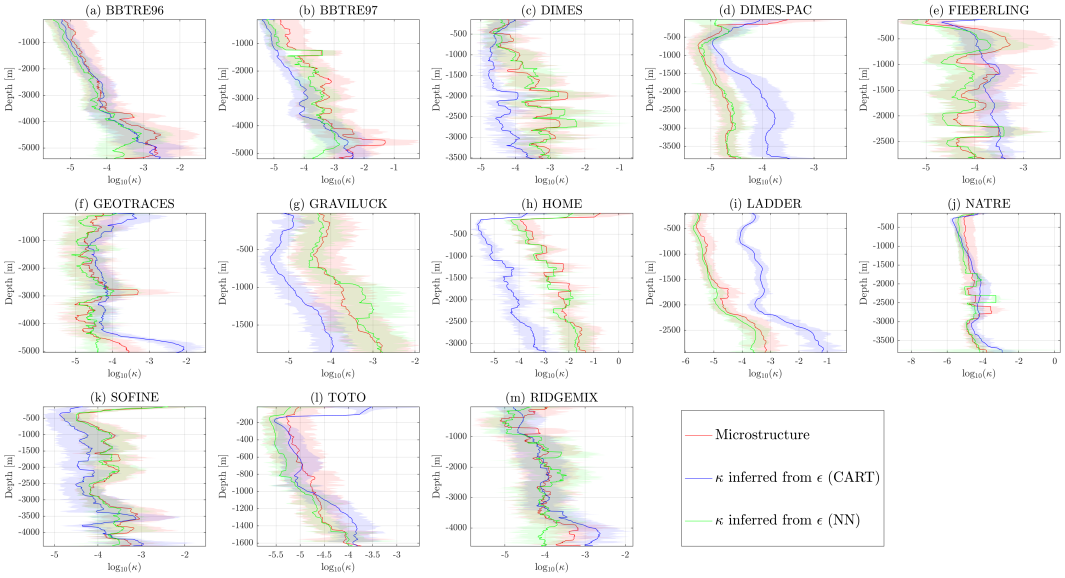

Figure 2. Same as Fig. 4 in the main text, but here the models are trained to predict ϵ and then κ is inferred from that prediction using Eq. (1) with $\Gamma = 0.2$.