AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters®

RESEARCH LETTER
10.1029/2022GL098039

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], and Colm-cille P. Caulfield[5]

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

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Mashayek,
mashayek@ic.ac.uk

Citation:

Mashayek, A., Reynard, N., Zhai, F., Srinivasan, K., Jelley, A., Naveira Garabato, A., & Caulfield, C. P. (2022). Deep ocean learning of small scale turbulence. *Geophysical Research Letters*, *49*, e2022GL098039. https://doi.org/10.1029/2022GL098039

Received 27 JAN 2022
Accepted 8 JUL 2022
Corrected 22 AUG 2022

This article was corrected on 22 AUG 2022. See the end of the full text for details.

Author Contributions:

Conceptualization: Ali Mashayek
Formal analysis: Ali Mashayek, Nick Reynard, Fangming Zhai, Kaushik Srinivasan
Funding acquisition: Ali Mashayek
Investigation: Ali Mashayek, Alberto Naveira Garabato, Colm-cille P. Caulfield
Methodology: Ali Mashayek, Nick Reynard, Fangming Zhai, Kaushik Srinivasan, Adam Jelley
Supervision: Ali Mashayek

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 ocean and 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 tracers 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.

1. 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 Supporting Information S1 for a visual summary, and Davis et al. (2019) for a review (Davis et al., 2019)—hereafter we refer to Supplementary Materials as SM). 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.

Validation: Nick Reynard, Fangming Zhai, Kaushik Srinivasan
Visualization: Ali Mashayek, Nick Reynard, Fangming Zhai
Writing – original draft: Ali Mashayek
Writing – review & editing: Nick Reynard, Kaushik Srinivasan, Adam Jelley, Alberto Naveira Garabato, Colmcille P. Caulfield

2. Physics of Turbulence

A key property of density-stratified ocean turbulence is the significantly enhanced rate at which it mixes density and tracers in the vertical (as compared to molecular diffusion; 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; 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 $N = \sqrt{-\left[g/\rho_0\right] \partial\rho/\partial z}$ is the buoyancy frequency. While Γ is known to be variable (Gregg et al., 2018; Mashayek & Peltier, 2013; Mashayek et al., 2017), 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*). We found that inclusion of both Z and *Hab* is crucial as they represent the distances 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 Figure S1 in Supporting Information S1). This allows for global inference of κ from observational surveys, thereby providing a route for application to regional high-resolution models and state estimates which assimilate data from such surveys (e.g., Forget et al., 2015; Verdy & Mazloff, 2017).

3. The Training Data Set

We employ a global data set of microstructure profiles compiled by the Climate Process Team on internal wave-driven ocean mixing (MacKinnon et al., 2017). Figure 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 Diapycnal and Isopycnal Mixing Experiment in Southern Ocean (DIMES) experiment is shown in panel *c* (more specifically, transect T1 in Figure 5a). Figure 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 data set 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 Figure 1 contain over 700 full-depth microstructure profiles, binned into 10 m vertical bins (amounting to $\sim 2 \times 10^5$ data points). The original post-processed ε data (prior to binning) has a vertical resolution of ~0.5 m; see text and Figure S2 in Supporting Information S1. The concurrent measurements of ε, T, S, Z in this data set allow for the construction of the aforementioned predictor list (i.e., the list of features used in training) used to predict daε 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 latest version of the global bathymetric map of Sandwell et al. (2014), and height above is then calculated by subtracting the sample depth from the local depth.

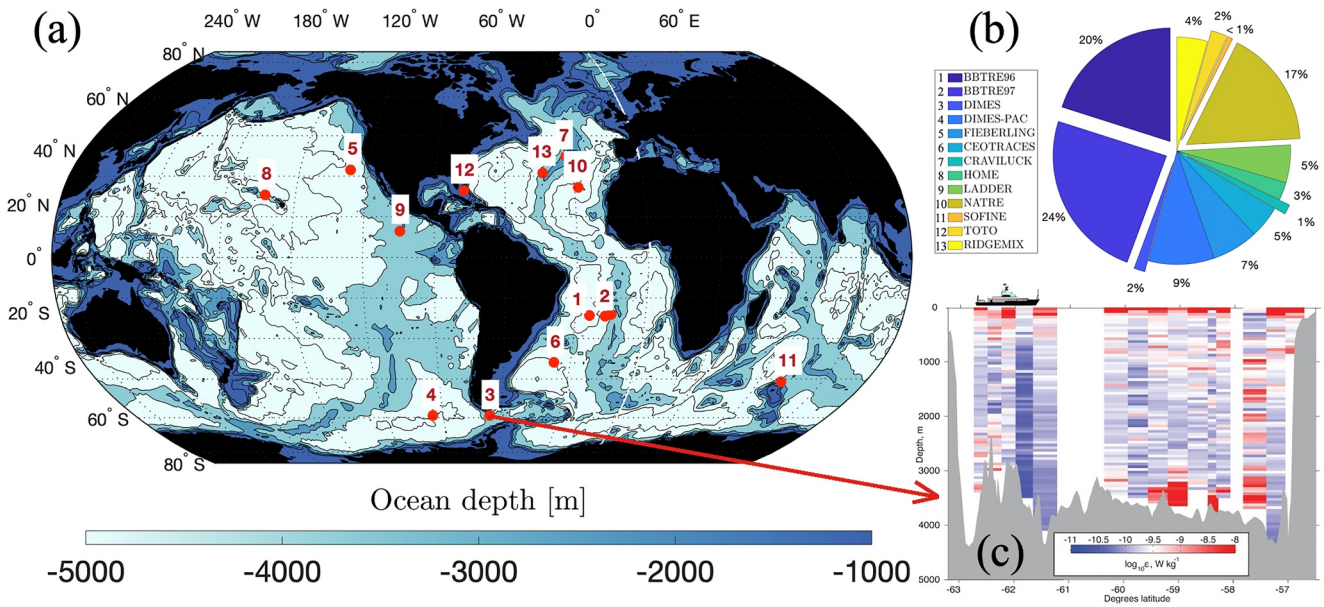

Figure 1. Direct turbulence measurements can be used to train machine-learning algorithms to predict turbulent mixing where direct measurements are not available. (a) Location of the field programs that include direct measurements of turbulence (specifically, turbulent kinetic energy dissipation rate 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 Supporting Information S1. 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 Diapycnal and Isopycnal Mixing Experiment in Southern Ocean (DIMES) experiment (transect T1 in Figure 5); from Sheen et al. (2013).

4. Machine Learning Models

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

4.1. Classification and Regression Trees (CART)

We employ CART, one of the most common machine learning predictive models (Breiman et al., 1984; Wu et al., 2008). The method uses a decision tree to connect observations of parameters of interest (represented in the branches) to predictions of a 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 averaging over the predictions of the trees. (Breiman, 1996).

Figures 3a and 3b show the application of the bagging tree to the training microstructure data set. The model was trained based on 10 cross-validation k-folds of all data across 13 field experiments. This method involves splitting the data set into k equally sized groups, or "folds," and taking it in turn to use each group as the validation 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 data set, hence reducing bias when compared to other methods. The fits in Figures 3a and 3b are satisfactory, with a coefficient of determination (R^2) of 0.83 for $\log_{10}(\varepsilon)$ and 0.84 for $\log_{10}(\kappa)$. R^2 is a statistical metric quantifying how well the regression predictions approximate the real data, and so is a measure of the goodness of fit of a model. To analyze further the quality of the agreement between predictions and data, panels e, f display the cumulative contribution of various predictors to increases in R^2 and decreases in the mean squared error (MSE).

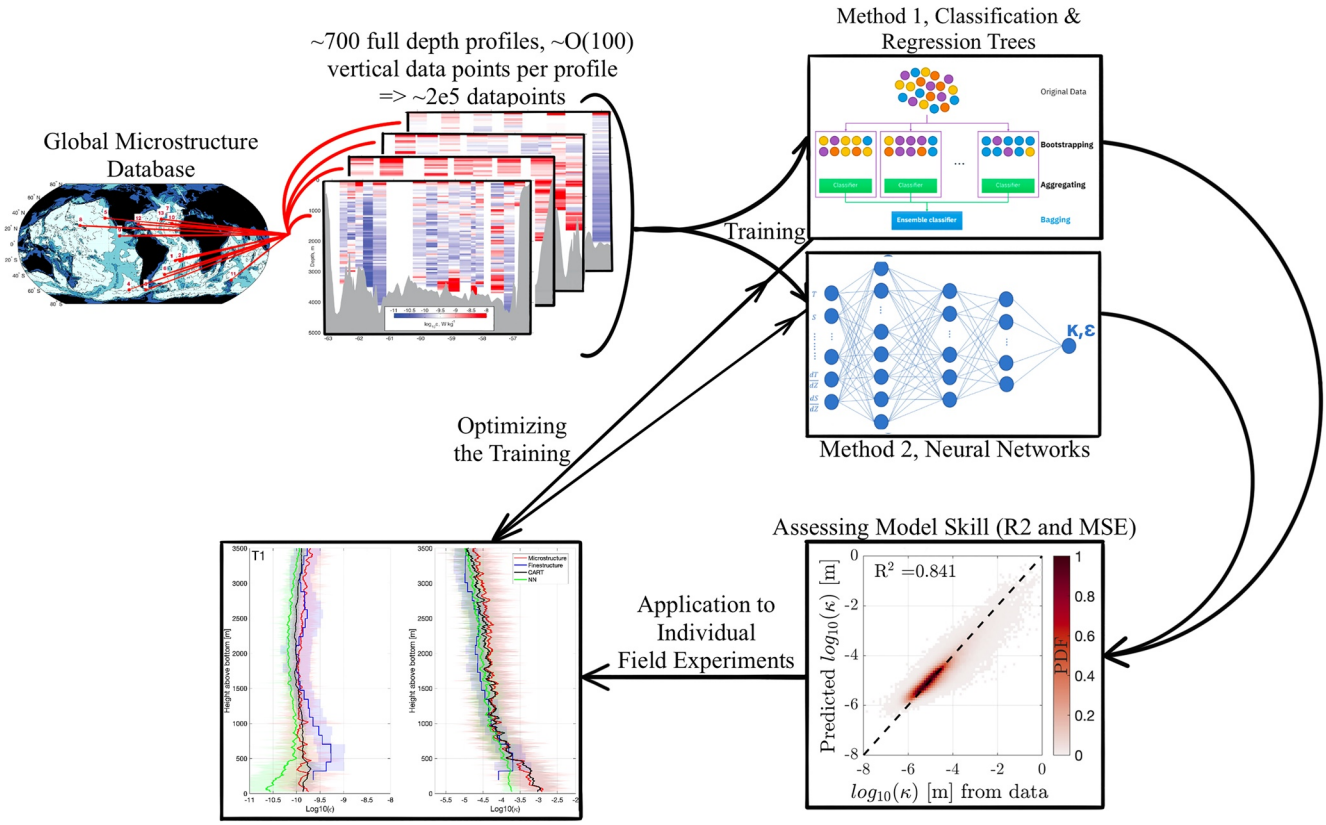

Figure 2. Two distinctly different machine-learning algorithms can successfully reproduce turbulent mixing estimates in agreement with microstructure data. A flowchart, illustrating the sequence of data assembly, training, model skill assessment, application to original data, verification, and fine-tuning.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

We consider T, S, their vertical gradients, Z, Hab, $\log_{10}(N^2)$, and latitude as our input predictors/feature. In calculation of N^2, we use the sorted background density (e.g., following Smyth et al., 2001) to avoid negative values in regions of unstable density stratification. This, along with inclusion of the actual T, S, and their gradients, provides information on both the background flow on which turbulence develops and the turbulence perturbations. 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. We also note that quantities like ε, κ, and N vary over orders of magnitude, and typically have log or log-skewed distributions (e.g., see Cael & Mashayek, 2021). Deviations from Gaussianity would need characterization of higher order moments or cumulants. Thus, from a training perspective, it is more sensible to consider the logarithm of these parameters as we do hereafter.

It is worth noting that we also tried another standard choice, namely the Least Squares Boost (LSBoost) algorithm, as an alternative learning algorithm, similar to CART but which additionally utilises gradient boosting (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, and thus is our method of choice. Finally, we note that application of a linear regression model to the data set proved entirely futile, as might be expected, since the relationship between the input features and turbulence is highly non-linear.

4.2. 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 in the broader

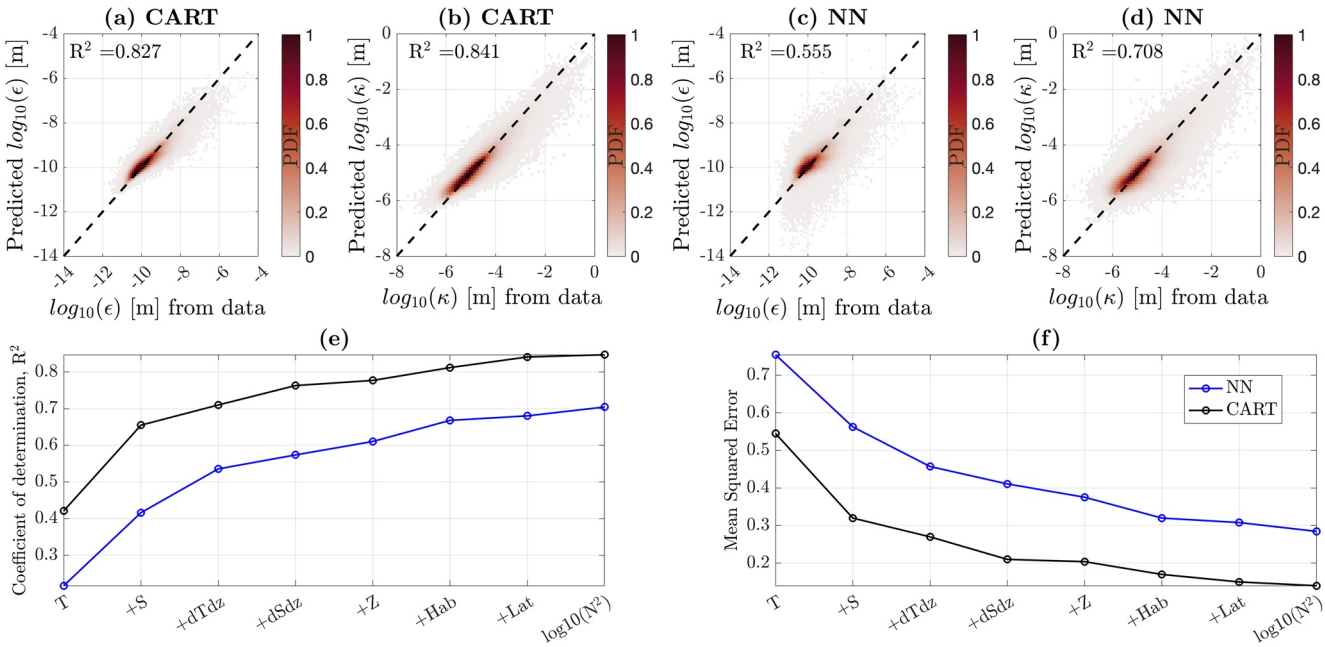

Figure 3. Machine learning can successfully fit the global microstructure data based on few predicting features. (a–d) Bivariate histograms (in the form of a probability density function, PDF) of predicted rate of dissipation of turbulent kinetic energy (ε; [m^2/s^3]) and turbulent diffusivity (κ; [m^2/s]) 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 and 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). All the data sets shown in Figure 1 are employed in this figure.

machine learning (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 (He et al., 2016; Ramachandran et al., 2017; Veit et al., 2016). 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 artificial intelligence/ML (AI/ML) in general (Drozdzal et al., 2016; Vaswani et al., 2017) and hence residual layers are correspondingly ubiquitous features of most modern NNs. In this work, 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, reducing the issue of overfitting, even in low data regimes. A 10-fold cross-validation algorithm is used to ensure coverage of the entire data set using the trained NN models. The SM contain more information on the resnet-FNN model architecture, as well as details of the training and optimization procedure.

Figures 3c and 3d show that the deep neural network is also skillful in predicting both $\log_{10}(\varepsilon)$ and $\log_{10}(\kappa)$. 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 Figure 3 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.

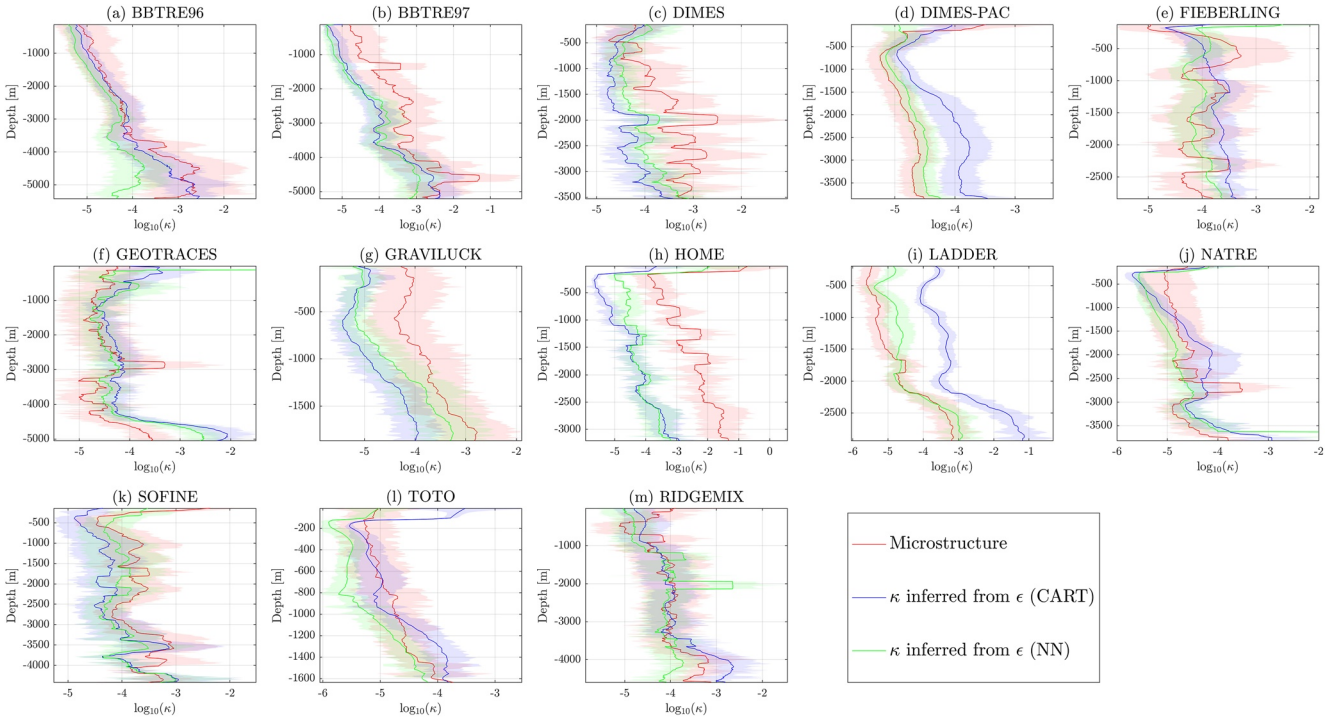

Figure 4. Predictions for individual field programs also show promise. Comparison of the predictions of the machine learning algorithms (Classification And Regression Tree (CART) and neural network (NN)) to each of the 13 field programs introduced in Figure 1. For each case, the solid lines represent 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. Averaged profiles are constructed as averages of $\log_{10}(\kappa)$. For each experiment, the models were trained based on the data from all other 12 experiments, excluding the data from the given experiment itself (i.e., out-of-distribution detection). For this plot, the models were trained to predict ε from which κ was inferred using Equation 1. A similar plot, showing qualitatively the same level of success, is included in SM, in which the models were trained to directly predict κ.

5. Application to Individual Data Sets

Figure 4 shows the results of separate analyses for each of the 13 different field programs listed in Figure 1. Importantly, the data from each experiment are excluded from training of the models before the models are applied to it (i.e., out-of-distribution detection). Both NN and CART show skills in predicting the patterns and, in some cases, the order of magnitude adequately, with their comparative success varying from case to case. For Figure 4, the models were trained to predict $\log_{10}(\varepsilon)$, from which $\log_{10}(\kappa)$ was inferred using Equation 1 with $\Gamma = 0.2$. We also repeated the exercise based on models trained to directly predict $\log_{10}(\kappa)$ (see Figure S3 in Supporting Information S1). Again, the comparative skill of direct versus indirect predictions varied from case to case. Understanding the reasons behind varying levels of success from case to case between the direct and indirect prediction of κ, between NN and CART, and between the upper and deeper parts of the water column, requires an in-depth analysis of the underlying phenomenology of turbulence, which would be overwhelming here given the number of experiments considered. However, such analysis, on a case by case basis, can form a natural extension of this work.

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 Figure 4, in light of its agnosticism toward the underlying physics, is appealing. To further highlight this point, in Figure 5 we assess the skill of CART and NN for the data sampled along three transects (shown in Figure 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 (Figures 5b–5g). Both NN and CART prove superior over the finestructure parameterization, particularly for κ (which is ultimately the parameter of interest). The superiority is even more

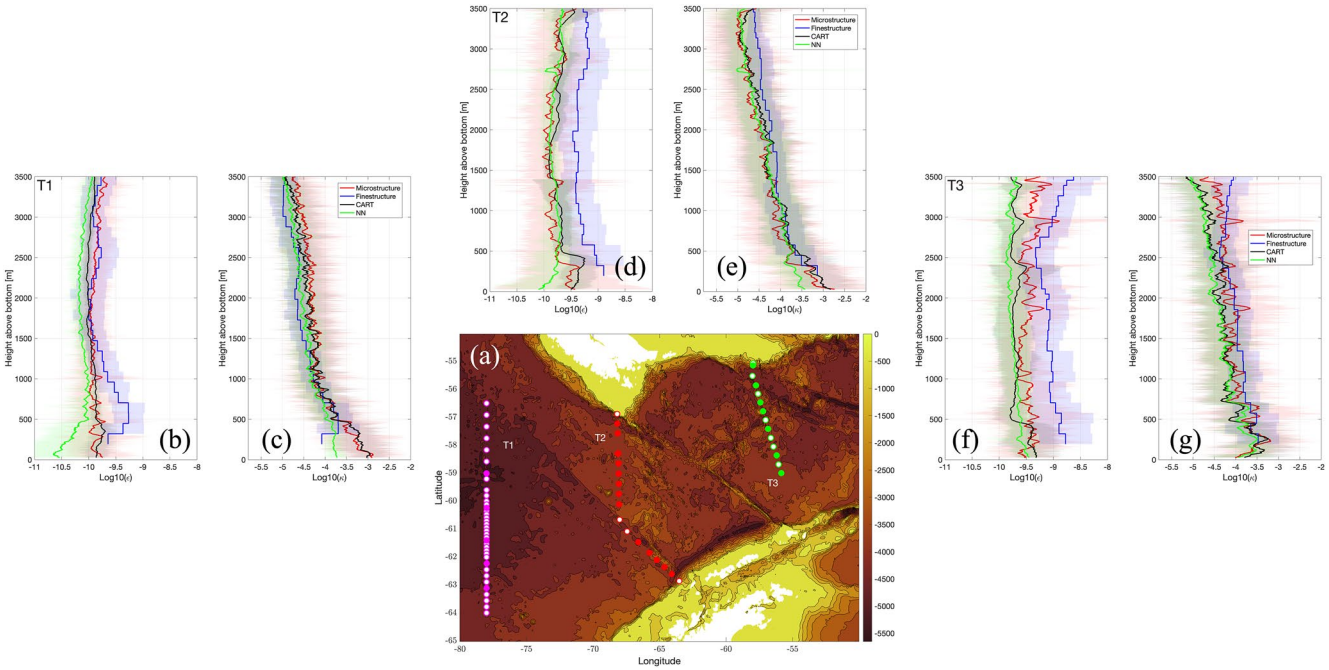

Figure 5. Machine learning competes well against physics-based parameterizations. Comparison of the predictions from machine learning against finestructure parameterization for three transects of the DIMES experiment. (a) The Drake Passage of the Southern Ocean, with the three cruise transects marked. Each circle represents a sampling. Transect T1 is the most western, plotted in magenta. Filled circles mark locations where microstructure data were taken, along with T, S, Z (from which finestructure-based estimates 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 Classification And Regression Tree (CART) and neural network (NN)) estimates of ε and κ for transect T1 (in b and c), transect T2 (in d and e), and transect T3 (in f and g). Two figures in SM show similar plots but for individual profiles along T2. 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).

pronounced when predictions for individual profiles are considered (see Figures S4 and S5 in Supporting Information S1). 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.

6. Discussion and 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, that is, 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 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 Figure 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, for example, 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 data sets. 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 data sets 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 toward 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. The machine learning algorithms developed here are applicable to data sets which preserve the "wiggles" that appear in T and S profiles due to turbulence. As discussed earlier (also see Figure S2 in Supporting Information S1), the efficiency of the algorithms diminishes with a decreased vertical resolution of the data. The resolution employed here (10 m) is perhaps a starting point. Increasing the resolution, while possible from the data availability perspective, can make training less efficient (the smoother the data, the easier the regression). Thus, exploring the optimal training resolution is a natural extension and (similar to a more in-depth analysis of comparative skills of NN and CART) will need to be done on a case-by-case basis. Given that the typical coarse vertical resolution of climate models does not allow for footprint of small-scale turbulence to appear in the model solutions, our machine learning algorithms are not directly applicable to such models. However, they can be excellent tools for assimilating data in state estimates, which typically have higher vertical resolutions and include direct or indirect information on spatio-temporal distributions of mixing and dissipation. Furthermore, our methodology can readily be applied to regional and process study models with sufficient spatio-temporal resolution. We finish by noting that one crucial information for diapycnal mixing parameterization, from a physical perspective, is velocity shear. While shear is rarely measured concurrently with microstructure data, it is available from model solutions. Thus, another extension of our analyses would be to train algorithms on a select number of direct observations, which include shear measurements (a much smaller subset of our training data). This would allow for shear to be added to the predictor list and will render the resulting machine learning algorithms better suited for application to operational ocean models.

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 name list of the experiments includes BBTRe, DIMES, FIEBERLING, GEOTRACERS, GRAVILUCK, HOME, LADDER, NATRE, SOFINE, TOTO, and RIDGEMIX. The AI algorithms are available at https://doi.org/10.5281/zenodo.6857164.

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 Figure 5, and Lois Baker and two anonymous reviewers for constructive comments. A.M. acknowledges support from National Environmental Research Council (NE/P018319/1), and N.R. was supported by the Centre for Doctoral Training in Sustainable Environmental Engineering, UK EPSRC funded (EP/L016826). The collection of these data took place after years of instrument development and hundreds of at-sea days, and would not have been possible without the hard work and skill of the Captain and crew of each research vessel.

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. https://doi.org/10.1007/bf00058655
Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *Wadsworth International Group*, *37*(15), 237–251.
Cael, B., & Mashayek, A. (2021). Log-skew-normality of ocean turbulence. *Physical Review Letters*, *126*(22), 224502. https://doi.org/10.1103/physrevlett.126.224502
Davis, R. E., Talley, L. D., Roemmich, D., Owens, W. B., Rudnick, D. L., Toole, J., et al. (2019). 100 years of progress in ocean observing systems. *Meteorological Monographs*, *59*(1), 3.1–3.46. https://doi.org/10.1175/amsmonographs-d-18-0014.1
de Lavergne, C., Falahat, S., Madec, G., Roquet, F., Nycander, J., & Vic, C. (2019). Toward global maps of internal tide energy sinks. *Ocean Modelling*, *137*, 52–75. https://doi.org/10.1016/j.ocemod.2019.03.010
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. https://doi.org/10.1007/978-3-319-46976-8_19
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. *Geoscientific Model Development*, *8*, 3071–3104. https://doi.org/10.5194/gmd-8-3071-2015
Garabato, A. N., & Meredith, M. (2022). Ocean mixing: Oceanography at a watershed. In *Ocean mixing* (pp. 1–4). Elsevier.
GEOTRACERS. (2019). Geotracers. Retrieved from https://www.geotraces.org/
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org
Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.
GO-SHIP. (2018). Go-ship. 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). Mixing efficiency in the ocean. *Annual Review of Marine Science*, *10*(1), 443–473. https://doi.org/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). https://doi.org/10.1109/cvpr.2016.90

Jackett, D. R., & McDougall, T. J. (1997). A neutral density variable for the world's oceans. *Journal of Physical Oceanography*, *27*(2), 237–263. https://doi.org/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., et al. (2017). Climate process team on internal wave-driven ocean mixing. *Bulletin of the American Meteorological Society*, *98*(11), 2429–2454. https://doi.org/10.1175/BAMS-D-16-0030.1

Mashayek, A., Caulfield, C. P., & Alford, M. H. (2021). Goldilocks mixing in oceanic shear-induced turbulent overturns. *Journal of Fluid Mechanics*, *928*, A1. https://doi.org/10.1017/jfm.2021.740

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. https://doi.org/10.1017/jfm.2013.176

Mashayek, A., Salehipour, H., Bouffard, D., Caulfield, C. P., Ferrari, R., Nikurashin, M., et al. (2017). Efficiency of turbulent mixing in the abyssal ocean circulation. *Geophysical Research Letters*, *44*(12), 6296–6306. https://doi.org/10.1002/2016gl072452

Middleton, L., Fine, E., MacKinnon, J., Alford, M., & Taylor, J. (2021). Estimating dissipation rates associated with double diffusion. *Geophysical Research Letters*, *48*(15), e2021GL092779. https://doi.org/10.1029/2021gl092779

Peltier, W. R., & Caulfield, C. P. (2003). Mixing efficiency in stratified shear flows. *Annual Review of Fluid Mechanics*, *35*, 135–167. https://doi.org/10.1146/annurev.fluid.35.101101.161144

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. https://doi.org/10.1002/2013jc008979

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

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. *Frontiers in Marine Science*, *6*(1), 439. https://doi.org/10.3389/fmars.2019.00439

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. https://doi.org/10.1126/science.1258213

Sheen, K. L., Brearley, J. A., Naveira Garabato, A. C., Smeed, D. A., Waterman, S., Ledwell, J. R., et al. (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. https://doi.org/10.1002/jgrc.20217

Shroyer, E. L., Nash, J. D., Waterhouse, A. F., & Moum, J. N. (2018). Measuring ocean turbulence. In *Observing the oceans in real time. Springer Oceanography* (pp. 99–122). Springer. https://doi.org/10.1007/978-3-319-66493-4_6

Smyth, W. D., Moum, J., & Caldwell, D. (2001). The efficiency of mixing in turbulent patches: Inferences from direct simulations and microstructure observations. *Journal of Physical Oceanography*, *31*, 1969–1992. https://doi.org/10.1175/1520-0485(2001)031<1969:teomit>2.0.co;2

Talley, L. D., Feely, R. A., Sloyan, B. M., Wanninkhof, R., Baringer, M. O., Bullister, J. L., et al. (2016). 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. https://doi.org/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., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. *Advances in Neural Information Processing Systems*, *29*, 550–558.

Verdy, A., & Mazloff, M. R. (2017). A data assimilating model for estimating Southern Ocean biogeochemistry. *JGR Oceans*, *122*(9), 6968–6988. https://doi.org/10.1002/2016JC012650

Waterhouse, A. F., MacKinnon, J. A., Nash, J. D., Alford, M. H., Kunze, E., Simmons, H. L., et al. (2014). Global patterns of diapycnal mixing from measurements of the turbulent dissipation rate. *Journal of Physical Oceanography*, *44*(7), 1854–1872. https://doi.org/10.1175/JPO-D-13-0104.1

Whalen, C., Talley, L., & MacKinnon, J. (2012). Spatial and temporal variability of global ocean mixing inferred from Argo profiles. *Geophysical Research Letters*, *39*(18). https://doi.org/10.1029/2012gl053196

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Erratum

In the originally published version of this article, author Colm-cille P. Caulfield's name was changed in error, and has now been restored to its original, correct spelling. The present version may be considered the authoritative version of record.