

Data-Driven Probabilistic Air-Sea Flux Parameterization

Jiarong Wu¹, Pavel Perezhogin¹, David John Gagne², Brandon Reichl³,
Aneesh C. Subramanian⁴, Elizabeth Thompson⁵, and Laure Zanna¹

¹Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

²National Center for Atmospheric Research, Boulder, CO, USA

³NOAA – Geophysical Fluids Dynamics Laboratory, Princeton, NJ, USA

⁴Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA

⁵NOAA Physical Sciences Lab, Boulder, CO, USA

Key Points:

- We propose a probabilistic air-sea turbulent momentum and heat flux algorithm based on neural networks trained on in-situ observations.
- Our algorithm quantifies the uncertainty (variability) around the mean prediction, while the mean is similar to existing bulk algorithms.
- Deterministic and stochastic tests on forced single-column upper-ocean model are performed, and the results highlight seasonal responses.

Abstract

Accurately quantifying air-sea fluxes is important for understanding air-sea interactions and improving coupled weather and climate systems. This study introduces a probabilistic framework to represent the highly variable nature of air-sea fluxes, which is missing in deterministic bulk algorithms. Assuming Gaussian distributions conditioned on the input variables, we use artificial neural networks and eddy-covariance measurement data to estimate the mean and variance by minimizing negative log-likelihood loss. The trained neural networks provide alternative mean flux estimates to existing bulk algorithms, and quantify the uncertainty around the mean estimates. Stochastic parameterization of air-sea turbulent fluxes can be constructed by sampling from the predicted distributions. Tests in a single-column forced upper-ocean model suggest that changes in flux algorithms influence sea surface temperature and mixed layer depth seasonally. The ensemble spread in stochastic runs is most pronounced during spring restratification.

1 Introduction

The atmosphere and ocean exchange mass, momentum, and heat across the air-sea interface. These fluxes influence ocean and atmosphere processes across a vast range of scales. However, quantifying air-sea fluxes is challenging, and there are still significant uncertainties (Cronin et al., 2019).

In this work, we focus on momentum flux (denoted as τ_x and τ_y) and turbulent heat flux (THF), which is a non-radiative heat flux consisting of sensible heat flux (Q_S) due to air-sea temperature difference and latent heat flux (Q_L) due to air-sea humidity difference. Direct in-situ observations of these turbulent fluxes from atmosphere to ocean rely on measuring the covariances of turbulent fluctuations:

$$\tau_x = -\rho_a \overline{u'w'}, \quad \tau_y = -\rho_a \overline{v'w'}, \quad Q_S = -\rho_a c_p \overline{w'T'}, \quad Q_L = -\rho_a L_e \overline{w'q'}. \quad (1)$$

Here u' , v' , w' , T' , and q' are fluctuations of three velocity components, potential temperature, and specific humidity; ρ_a is the air density; c_p is the specific heat capacity at constant pressure, and L_e is the latent heat of evaporation. The measurements are performed in the atmospheric surface layer, where these fluxes are assumed to be constant (Fairall et al., 1996).

The challenges of quantifying these fluxes lie in both observations and modeling. Due to the requirements of sophisticated instruments and careful quality control, such direct measurements are usually carried out on designated research cruises (Bradley & Fairall, 2006). Less equipped measuring platforms and remote sensing measure atmospheric and oceanic surface variables (wind speed, temperature, humidity, etc.) and rely on a bulk flux algorithm to compute the fluxes from the mean observed variables. The same algorithm is used in coupled weather and climate models to compute flux based on prognostic state variables of the oceanic and atmospheric surfaces.

The widely used flux algorithms are termed bulk algorithms since they use bulk quantities of the surface layer to model the surface fluxes. The momentum and turbulent heat fluxes are formulated as being proportional to the magnitude of wind speed $|U_a|$ and air-sea difference in velocity, temperature, and humidity

$$\tau_x = \rho_a C_D S(U_a - U_o), \quad Q_S = \rho_a c_p C_H S(T_a - T_o), \quad Q_L = \rho_a L_e C_E S(q_a - q_s), \quad (2)$$

where C_D , C_H , and C_E are the transfer coefficients for momentum, sensible, and latent heat, respectively. S is the scalar wind speed relative to the ocean surface (subject to gustiness correction). Here, τ_x is aligned with the surface wind and the cross-wind component τ_y is assumed zero. The transfer coefficients are calculated based on the Monin-Obukhov similarity theory, where certain parameters (stability function and roughness length) are empirically determined from observations; see a complete description in, e.g., Fairall et al. (2003) for COARE algorithm.

Currently, different bulk algorithms exist fitted to different sets of observations (Brunke et al., 2003; Biri et al., 2023). There are also varying levels of simplifications and empirical corrections (e.g., cool-skin warm-layer or gustiness corrections). This is a source of uncertainty for flux estimation in general circulation models (GCMs) and flux products. Sensitivity studies have shown that changing bulk algorithms can considerably affect atmospheric dynamics, e.g. in terms of Madden Julian Oscillation (Hsu et al., 2022), precipitation (Harrop et al., 2018), general circulation (Polichtchouk & Shepherd, 2016), and oceanic state, e.g. in terms of sea surface temperature (Bonino et al., 2022). Flux products also suffer from uncertainty in bulk algorithms, in addition to uncertainty in bulk inputs, as discussed in Yu (2019).

Because of their prohibitively high computational costs, high-fidelity numerical simulations are not yet widely used to estimate air-sea turbulent fluxes across diverse conditions. As a result, eddy-covariance (EC) measurements remain our best “ground-truth” for calibrating bulk algorithms, despite their sparsity and intrinsic measurement uncertainty (Gleckler & Weare, 1997). We use a quality-controlled research ship cruise dataset provided by NOAA Physical Sciences Lab (PSL) to develop an alternative data-driven flux algorithm using artificial neural networks (ANNs). Recent advances in machine learning methods have led to their applications to surface flux parameterization, see e.g. McCandless et al. (2022); Leufen and Schädler (2019) for land-atmosphere fluxes, Cummins et al. (2024) for polar regions, and Zhou et al. (2024) for air-sea heat fluxes. Earlier attempts to replace iterative bulk algorithms with computationally efficient neural networks date back to Bourras et al. (2007).

Another potential drawback of deterministic bulk algorithms is the lack of variability. Bulk algorithms are simplified representations of the underlying dynamical processes and, at best, represent a statistically averaged value of fluxes given the observables. There is a significant spread of observed EC flux data around the prediction of bulk algorithms. Such deviations from the mean flux values may be crucial for modeling processes on small and fast scales (Nuijens et al., 2024). They can also have a rectifying effect for the large-scale dynamics in nonlinear GCMs. More complex formulas such as sea-state-dependent parameterization (Edson et al., 2013; Bouin et al., 2024) have been proposed that incorporate surface wave physics (one of the sources of additional variability) but have not yet been fully validated and widely adopted.

In this study, we aim to quantify the uncertainty (variability) around deterministic flux algorithms from EC data, while taking an agnostic perspective on the source of uncertainty. This is achieved by developing a probabilistic air-sea flux model based on conditional parametric probability distributions. This class of parametric models uses a finite number of parameters to represent the underlying probability distribution (Nix & Weigend, 1994; Barnes et al., 2021) and has been successfully applied to geophysical problems (Guillaumin & Zanna, 2021; Schreck et al., 2024). In most of the paper, we use the terms “uncertainty” and “variability” interchangeably to refer to the spread in observed data, but its attribution will be discussed in the end.

An estimate of uncertainty around deterministic bulk algorithms will benefit uncertainty quantification in downstream applications such as flux product development, and the development and testing of stochastic air-sea flux parameterization in weather and climate models. Stochastic parameterizations have been shown to improve the representation of variability and, in some cases, reduce bias in the mean state (Williams, 2012; Berner et al., 2017). In particular, air-sea fluxes are expected to play a significant role in the upper ocean. However, the sensitivity of the upper ocean states to uncertainty in the flux algorithm is not yet well quantified. We implement our probabilistic flux algorithm in the single-column General Ocean Turbulence Model (GOTM, Umlauf and Burchard (2005)) to study the effects on the upper ocean states.

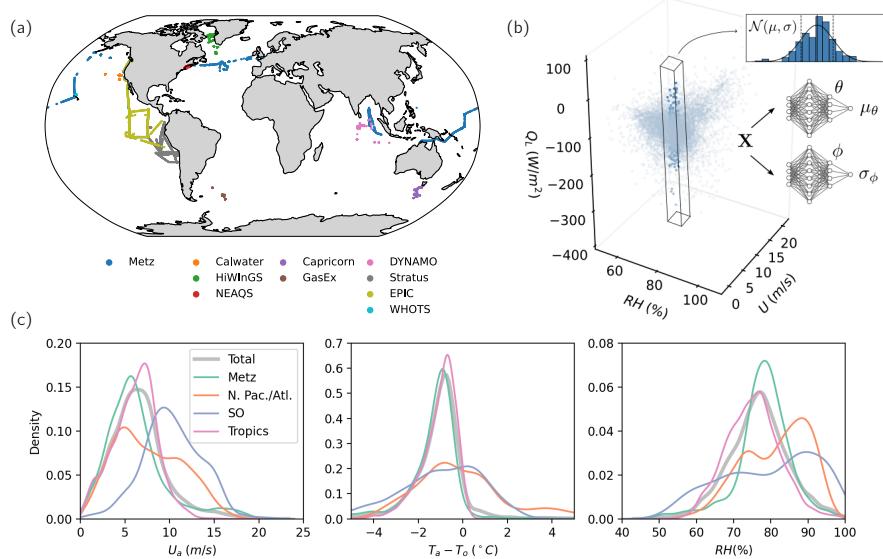


Figure 1. (a) Ship trajectories of the various cruises in the NOAA PSL dataset. (b) An illustration of the ANN-based conditional Gaussian probabilistic model. Note that we are visualizing only two of the input space dimensions, for the purpose of showing the concept of a conditional Gaussian distribution. (c) Distributions of input variables for different subsets of data.

The paper is organized as follows: in Section 2, we describe the dataset, the mathematical model, and the training procedure of ANN; in Section 3, we evaluate the ANN predictions compared to the baseline bulk algorithm and examine aspects of the learned flux model; in Section 4, the results of both deterministic and stochastic tests in GOTM are reported. We conclude by discussing the implications of our new probabilistic approach for air-sea flux modeling.

2 Data and model

2.1 Direct in-situ covariance measurements

The air-sea flux observation dataset we use is collected by decades of research cruises conducted by NOAA PSL. There are about 10,000 samples after quality control, and all are ship-borne, hourly-averaged covariance measurements according to Equation 1. The trajectories of the various cruises are shown in Figure 1(a). Among the labeled cruises, Metz is a compilation of multiple earlier cruises in the 1990s; WHOTS, EPIC, DYNAMO are conducted in the tropics and Stratus in the subtropics; Calwater, NEAQS, HiWInGS in mid/high latitude of the northern hemisphere, and Capricorn, GasEX in the southern ocean. Overall, this dataset covers various geographical locations, although a disproportional amount of data is collected in the tropics (more than 50%). We note that analyzing the measurements taken along transects as time series will likely reveal additional embedded information, but in this study, we contend with treating them as independent samples of the underlying conditional distribution.

2.2 Probabilistic model

We consider observed bulk variables \mathbf{X} (such as wind speed, humidity, etc.), unobserved variables \mathbf{Z} (for example sea state, vertical wind profile, etc.), and targeted flux

outputs $\mathbf{Y} = (\tau_x, \tau_y, Q_S, Q_L)$. Due to the unobserved variables \mathbf{Z} , the relation $\mathbf{Y} = \mathbf{F}_{model}(\mathbf{X})$ is not deterministic even when the physical laws governing the fluxes $\mathbf{Y} = \mathbf{F}_{true}(\mathbf{X}, \mathbf{Z})$ are deterministic. Using the proposed probabilistic model, we account for all uncertainties in the data, and further attributions are discussed in Section 5.

We assume that each flux component y of \mathbf{Y} follows a uni-variate conditional Gaussian distribution with mean $\mu(\mathbf{X})$ and standard deviation (std) $\sigma(\mathbf{X})$:

$$y \sim \mathcal{N}(\mu(\mathbf{X}), \sigma^2(\mathbf{X})). \quad (3)$$

In other words, $\mu(\mathbf{X})$ is our best unbiased estimation of the flux component y given \mathbf{X} , and the errors are distributed according to $\mathcal{N}(0, \sigma^2(\mathbf{X}))$ when conditioned on \mathbf{X} . This idea of conditional Gaussian distribution is demonstrated in Figure 1(b) with the latent heat flux data.

Having chosen the parametric distribution, we aim to learn the parameters $\mu_\theta(\mathbf{X})$ and $\sigma_\phi^2(\mathbf{X})$ with the given data. Here θ and ϕ denote the learnable parameters in the data-driven models for μ and σ^2 . The optimization procedure is based on computing the probability as a function of parameters (i.e., likelihood) of observing a sample value y given the sample input features \mathbf{x} :

$$p(y|\mathbf{x}, \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma_\phi^2(\mathbf{x})}} \exp \left[-\frac{(y - \mu_\theta(\mathbf{x}))^2}{2\sigma_\phi^2(\mathbf{x})} \right]. \quad (4)$$

The optimal model parameters θ and ϕ minimize this negative log-likelihood loss, computed on the training dataset summed over a total of N_{sample} (Nix & Weigend, 1994):

$$L_{nll}(\theta, \phi) = \sum_{m=1}^{N_{sample}} \frac{1}{2} \left[\log(\sigma_\phi^2(\mathbf{x}_m)) + \frac{(y_m - \mu_\theta(\mathbf{x}_m))^2}{\sigma_\phi^2(\mathbf{x}_m)} \right] + \text{const..} \quad (5)$$

2.3 ANN architecture and training

The mathematical framework described in Section 2.2 is general and applies to any data-driven model that approximates parametric distributions. In particular, as shown in Figure 1(b), we use two ANNs to represent the mean μ_θ and the variance σ_ϕ^2 for each flux component. In this case, θ and ϕ are the weights and biases of the ANNs. The non-linear activation function of the hidden layers is the Sigmoid function. The positivity of the predicted variance σ_ϕ^2 is ensured with the exponential activation function in the last ANN layer.

The input variables are selected based on a balance between expressibility and risk of over-fitting. After testing different numbers of inputs and their combinations, we have chosen five inputs

$$\mathbf{X} = (U_a, T_a, T_o, RH, p_a) \quad (6)$$

which are wind speed (measured between 16 to 21 m), atmospheric surface temperature (measured between 12 to 19.5 m), sea surface temperature (measured by sea snake at 0.05 m depth), relative humidity (measured between 12 to 19.5 m), and atmospheric pressure. Using relative humidity instead of specific humidity gives better behaviors during training, although the two can be converted given atmospheric temperature and pressure. We do not include the heights at which the meteorological variables are measured, although they are used as inputs to existing bulk algorithms.

The training of ANNs is described in detail in supplementary information. Briefly summarized here, we first train the networks on mean square error loss and then on negative log-likelihood loss (Equation 5). This two-step procedure allows the mean ANN to capture more variability in the data. The small data limit and the distribution shift

shown in Figure 1(c) are challenging for data-driven models, which we overcome through input feature selection, model design, training strategies (such as early stopping), and cross-validation.

Having learned the parametric distribution of the fluxes for the given inputs, predictions can be made in two ways: using the mean μ_θ for deterministic predictions or sampling from the distribution $\mathcal{N}(\mu_\theta, \sigma_\phi^2)$ for stochastic predictions. In Section 3.1, we evaluate the statistical scores of the deterministic predictions. In Section 3.2, we analyze the predicted dependence of μ_θ and σ_ϕ^2 on inputs. In Section 4, numerical experiments using both deterministic and stochastic heat fluxes are discussed. For simplicity, we denote the mean predictions as $(\mu_{\tau_x}, \mu_{\tau_y}, \mu_{Q_S}, \mu_{Q_L})$ and variance predictions as $(\sigma_{\tau_x}^2, \sigma_{\tau_y}^2, \sigma_{Q_S}^2, \sigma_{Q_L}^2)$, omitting subscript θ and ϕ .

3 Evaluation of the ANN-based probabilistic air-sea flux model

3.1 Statistical scores of the deterministic predictions

We evaluate the statistical scores of the flux predictions in terms of the root-mean-square-error (RMSE) and the coefficient of determination (R^2):

$$\text{RMSE}(\hat{y}, y) = (\mathbb{E}[(\hat{y} - y)^2])^{1/2}, \quad (7)$$

$$R^2(\hat{y}, y) = 1 - \mathbb{E}[(\hat{y} - y)^2]/\text{Var}[y]. \quad (8)$$

Here \hat{y} is the algorithm prediction and y is the truth, in our case the EC measurement $(\tau_{x,c}, \tau_{y,c}, Q_{S,c}, Q_{L,c})$. We evaluate these metrics for the ANN-based deterministic prediction $(\mu_{\tau_x}, \mu_{\tau_y}, \mu_{Q_S}, \mu_{Q_L})$ compared to a baseline bulk algorithm $(\tau_{x,b}, 0, Q_{S,b}, Q_{L,b})$. In particular, we choose COARE 3.0 (Fairall et al., 2003), and the scores are similar for the more updated version COARE 3.6. It is worth mentioning that COARE algorithms are fitted to the same EC data we use but with additional cruises.

Figure 2(a) shows the flux predictions plotted against the EC measurements. ANN-based deterministic predictions have similar but slightly higher R^2 than COARE for all three fluxes, see legends in Figure 2(a), with only a subset of the input variables. For the cross-wind momentum flux τ_y (not shown here), the ANN-based deterministic prediction has little predictive skills, as there is no input variable to indicate the sign of cross-wind stress. Without additional input features (such as surface wave information), it is reasonable to accept zero prediction as is in bulk algorithms. However, we can quantify the variability around the zero prediction with the current framework.

The statistical scores of both ANN and COARE vary considerably between different turbulent fluxes. The sensible heat flux Q_S is the hardest to predict for both based on its lowest R^2 values, while the latent heat flux Q_L gives the largest magnitude of error in terms of RMSE. Another observation is that scores of both ANN and COARE also vary considerably between geographical locations. Figure 2(b) shows R^2 of three fluxes evaluated on different subsets of cruises grouped roughly by geographical locations (same as in Figure 1). This is because each geographical subset has substantially different distributions of the input variable, as shown in Figure 1(c). In particular, R^2 for sensible heat flux Q_S is very low in the Tropics, only around 0.15. Overall, the fluctuation of predictive scores over different regions is shared between ANN and COARE. The biggest improvement seems to be for latent heat flux in some regions (up to about 0.2 increase in R^2). Supplementary table S1 lists RMSE and R^2 evaluated on the full dataset and on different geographical subsets.

3.2 Structure of the mean and variance predicted by ANNs

In addition to the statistical scores, we further evaluate the ANN predictions on a reduced-dimension uniform grid to probe and interpret the ANN flux model. This is

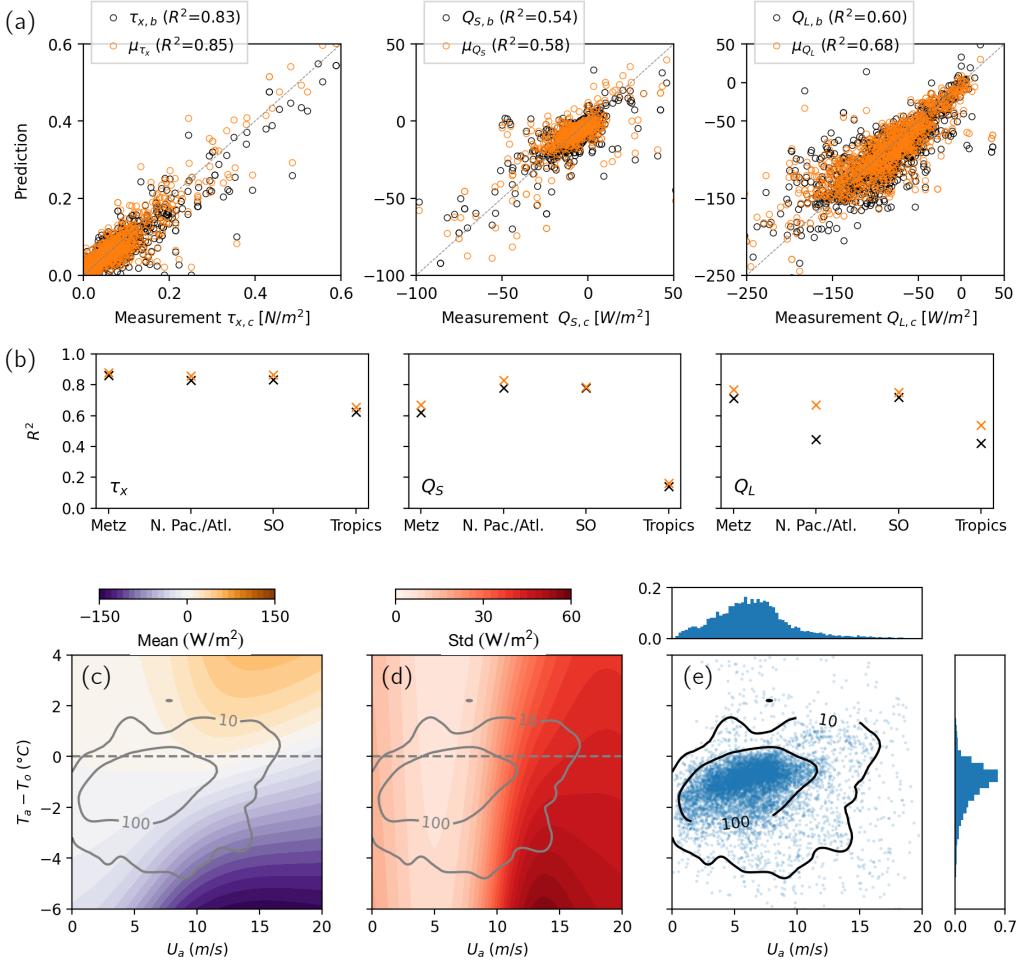


Figure 2. (a) ANN-based deterministic predictions (orange dots) and bulk algorithm predictions (black dots) plotted against measured fluxes. We are only visualizing 10% of total samples. (b) R^2 of ANN (orange) and bulk algorithm (black), evaluated on different geographical subsets. (c) ANN prediction of the mean of Q_s on a uniform input grid. The dashed gray lines mark zero temperature difference $T_a = T_o$. (d) ANN prediction of the std of Q_s for the same grid. (e) Scatter plot of all data points and their marginal distribution. Density estimated using Gaussian kernel density estimation provides the gray contour lines that indicate the available samples per unit ΔU and $\Delta(T_a - T_o)$. The same contour lines are overlaid on (c) and (d).

especially important for the std σ as a way to understand how the predicted std depends on input variables. Figure 2(c) and (d) illustrate this using sensible heat flux as an example. Among the five inputs, we fix sea surface temperature $T_o = 10$ $^{\circ}$ C, relative humidity $RH = 80\%$, and sea level pressure $p_a = 1010$ hPa, while varying wind speed U_a and temperature difference $T_a - T_o$. The plot represents a 2D slice of the 5D input space with typical values for the other variables.

Figure 2(c) shows the prediction of Q_s mean by ANN, which is smooth and asymmetric around $T_a - T_o = 0$. Unlike bulk algorithms that assume $Q_{s,b} \propto S(T_a - T_o)$, the structure of ANN is more flexible, allowing non-zero flux values even when $T_a = T_o$. A similar figure for COARE is shown in supplementary figure S1. We did not impose

any constraints on the sign of Q_S , contrary to Zhou et al. (2024) which used a penalty in loss function to ensure matching signs of $T_a - T_o$ and Q_S . In our experience, such a penalty is not necessary for a good statistical fit. Furthermore, there is insufficient evidence for a strictly down-gradient assumption, since locally counter-gradient fluxes are possible (Blay-Carreras et al., 2014; Deardorff, 1972) and systematic bias in measurements may exist.

Figure 2(d) shows the prediction of std of Q_S for the same grid. Uncertainty increases with wind speed, which is expected as residuals should scale with flux magnitude to some extent. A small but finite uncertainty persists at very low wind speed, likely due to measurement challenges for small fluxes. For a given wind speed, the uncertainty is consistently higher for $T_a < T_o$, corresponding to unstable boundary layer conditions. As shown in Figure 2(e), data are sparse for wind speeds above 15 m/s and large temperature differences. While the ANN seems to extrapolate reasonably well, it is worth exploring methods to further constrain it where training samples are not available. We note that the same issue applies to existing bulk algorithms as well.

4 Tests: surface flux forced upper ocean mixing

As a preliminary test for the proposed probabilistic air-sea flux model, we choose to implement it in the single column model GOTM (Umlauf & Burchard, 2005), which provides a controlled environment with no complex interplay with other nonlinear processes in GCMs. The governing equations are a set of 1D diffusion-type equations

$$\partial_t u = -\partial_z \overline{w' u'} + fv, \quad \partial_t v = -\partial_z \overline{w' v'} - fu, \quad \partial_t T = -\partial_z \overline{w' T'}, \quad \partial_t S = -\partial_z \overline{w' S'}, \quad (9)$$

where the prognostic variables are the horizontal velocity components u and v (subject to Coriolis force), temperature T , and salinity S . The turbulent fluxes $\overline{w' u'}$, $\overline{w' v'}$, $\overline{w' T'}$, $\overline{w' S'}$ are not resolved but are instead parameterized by vertical mixing schemes. In particular, we test our air-sea flux algorithms in combination with two ocean surface boundary layer (OSBL) mixing parameterizations: the K-profile-parameterization (KPP) scheme (Large et al., 1994) and a more sophisticated second-order closure $k-\epsilon$ scheme (Umlauf & Burchard, 2005), both used in GCMs but typically under different resolutions.

Air-sea momentum and heat fluxes serve as boundary conditions to Equation 9. In addition, the surface fluxes can affect the mixing parameterization. In the KPP scheme, a change in heat flux (therefore buoyancy flux), affects the OSBL depth through the Richardson number criterion, and the sign of buoyancy flux determines the shape function (Large et al., 1994). In the $k - \epsilon$ scheme, the production terms in the turbulence kinetic energy (TKE) equation are affected by surface fluxes.

4.1 Heat flux time series at OWS Papa

We discuss the flux time series first before describing the forced experiments. Figure 3(a) shows the time series of the ANN-predicted (deterministic) THF $Q = \mu_{Q_S} + \mu_{Q_L}$, in comparison to COARE 3.0. The differences between the ANN and COARE predictions are seasonal, because the input distributions vary across seasons. We demonstrate this by zooming into two months, June and October. In June, THF is low, and predictions align closely, while October shows larger discrepancies. The statistical significance of seasonality in flux difference is further demonstrated in Figure 4(c). We also plot two bulk algorithms - COARE 3.0 and NCAR (Large & Yeager, 2009) - to show that ANN deviations exceed those between different bulk algorithms. Within a period of high THF around Oct 20, ANN predicts the least extreme negative values while NCAR predicts the most extreme negative value, reflecting all three algorithms' divergence at high wind speed, where few measurements exist to constrain them.

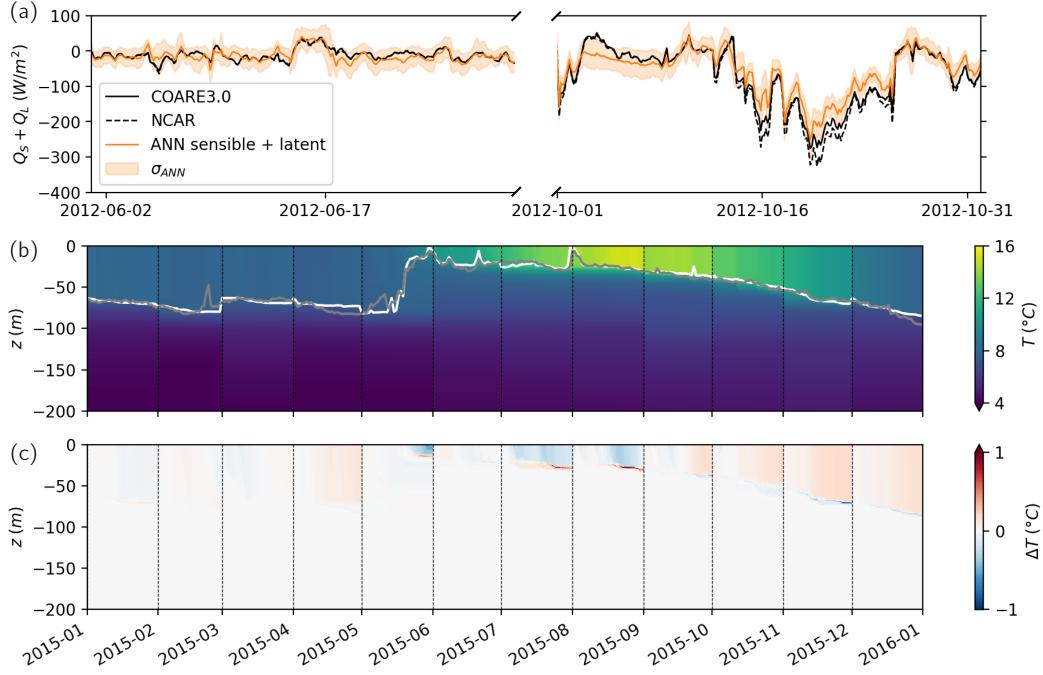


Figure 3. (a) Time series of THF for June and October. Green lines: heat flux computed using bulk algorithms, both COARE 3.0 (solid) and NCAR (dashed). Orange line: heat flux computed using ANN. Orange shade: $\pm\sigma$ for ANN flux. (b) Hovmöller diagram of temperature vertical profile over time, in the ANN flux forced case (k - ϵ scheme 10-minute time stepping). Monthly restart is shown with the black dotted lines. (c) Difference of temperature profiles $\Delta T(z, t) = T_{\text{ANN}}(z, t) - T_{\text{Bulk}}(z, t)$.

Figure 3(a) also shows the ANN-predicted $\pm 1\sigma_Q$ for THF in orange shading. We assume no covariance between the sensible and latent heat flux residuals, therefore, the uncertainty of their sum simply follows $\sigma_Q^2 = \sigma_{Q_s}^2 + \sigma_{Q_L}^2$. Importantly, the predicted σ_Q is state dependent, so the uncertainty envelope changes over time. It is also interesting to note that the bulk algorithm predicted THF is generally within $\pm\sigma_Q$.

There are two parts of the ANN-based probabilistic model to test. First, we want to evaluate the effects of changing the flux algorithm from COARE 3.0 to ANN-predicted mean in deterministic runs. Second, we want to propagate the uncertainty in fluxes through the single-column model and examine the uncertainty on the modeled upper ocean state by performing ensemble runs using stochastically perturbed fluxes. Since the uncertainty in THF is much larger than that in momentum flux, we change only the THF while using the same momentum flux as the control run. Details of the numerical experiments are in the supplementary information. There is a net heat flux imbalance due to ignoring horizontal advection, which we mitigate by restarting the simulation with the observed profiles monthly. For the same reason, our analysis emphasizes inter-comparisons between simulations rather than direct validation against observations.

4.2 Effects of changing deterministic heat flux forcing

Figure 3(b) shows a Hovmöller diagram of the typical annual cycle of temperature profiles. We focus on two characteristic quantities: sea surface temperature (SST) and mixed layer depth (MLD). MLD here is defined based on a potential density threshold

of 0.1 kg/m^3 compared to the surface value. The white and gray lines represent the simulated ($k - \epsilon$ scheme) and observed MLD, respectively, showing that the $k - \epsilon$ scheme reasonably tracks MLD evolution, at least given the regular restart with observed profiles. For instance, Figure 3(c) shows the difference in temperature profiles between ANN and COARE 3.0 flux forced runs. Such responses are summarized in Figure 4.

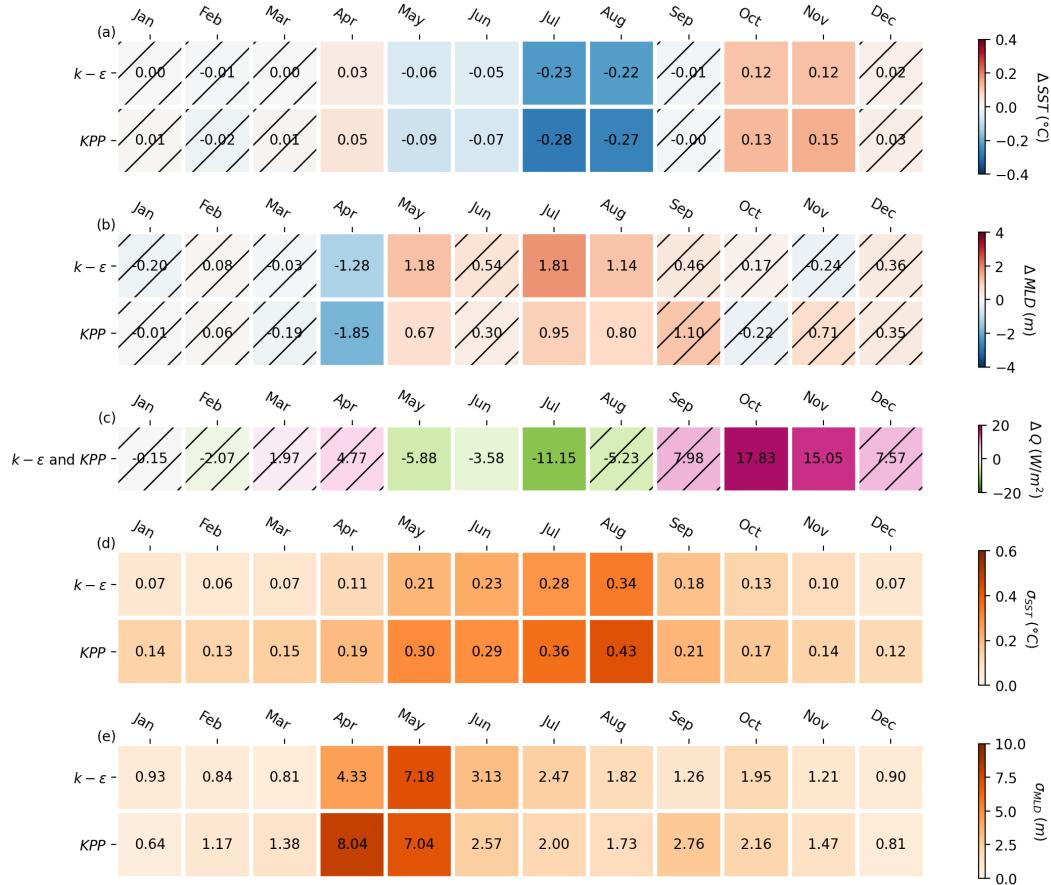


Figure 4. Difference in monthly-averaged SST (a) and MLD (b) between ANN flux runs and COARE flux runs at OWS Papa ensemble-averaged over 2011, 2012, 2015, and 2016. Trends with inconsistent signs over years are masked with hatching. The monthly-averaged THF discrepancies are shown in (c). Spread in MLD (d) and SST (e) induced by stochastic perturbation are calculated based on 20 ensemble members.

Figure 4(a) and (b) show the monthly-averaged change in SST and MLD, respectively. Again the effects of ANN forcing seem highly seasonal. In summer (July to August), it deepens the mixed layer and cools the sea surface. During spring restratification (April to May), ANN forcing tends to cause earlier restratification initially but a deeper mixed layer further into the summer. This is shown by negative ΔMLD in April and positive ΔMLD in May. These effects are consistent across $k - \epsilon$ and KPP runs, suggesting insensitivity to OSBL parameterization. In Figure 4(c), we plot the monthly-averaged difference in total heat flux (ANN minus COARE). Overall, there is more THF out of the ocean in summer and less in fall to early winter, roughly corresponding to the change we see in SST and MLD. However, the most salient net heat flux difference and SST/MLD changes are not always aligned in time, indicating more complex dynamical factors at play than merely a heat budget balance.

In summary, switching from COARE 3.0 to ANN can result in differences of up to 2 meters in monthly-averaged MLD and 0.2 degree in monthly-averaged SST. The instantaneous response is larger, although not shown here. For context, the difference in monthly-averaged MLD between KPP and $k - \epsilon$ mixing schemes can reach 16 meters, which is much larger in magnitude (see supplementary figure S2). However, changing OSBL parameterization typically induces a consistent year-round trend — KPP leads to a shallower mixed layer and warmer SST — whereas changing flux algorithm exhibits seasonality. This suggests that improving air-sea flux algorithms can potentially help correct seasonal MLD biases, which OSBL refinements alone may not fully address.

4.3 Stochastic ensemble runs

Given the ANN-predicted time series of mean and std of THF shown in Figure 3(a), we created an ensemble of stochastically perturbed fluxes. This ensemble of forced simulations provides an estimate of the variability in SST and MLD induced by the variability of THF. The stochastic perturbation is expected to stem from certain unobserved processes with characteristic correlation time scales. Therefore, instead of white noise, we use temporally correlated noise ϵ generated through a discrete autoregressive model of order one - AR(1), with a correlation time of 60 hours estimated from the EC data (see supplementary information for details).

From the 20 ensemble member runs conducted, we do not see any systematic drift induced by the stochastic forcing. In other words, the ensemble mean of stochastic runs is very similar to the deterministic run (see supplementary figure S3). However, there is a certain level of ensemble spread, which is summarized in Figure 4(d) and (e). The largest spread in MLD can be seen during the months of April and May, since stochasticity can significantly affect the onset of restratification. The largest spread in SST is seen around August. The reason may be that SST is sensitive to surface heating and cooling because the mixed layer is shallow in the summer. As in the deterministic runs, the effects seen in stochastic runs are not particularly sensitive to different OSBL parameterizations.

5 Conclusion and discussion

We propose an ANN-based probabilistic model for turbulent air-sea fluxes. State-of-the-art bulk algorithms represent the statistical mean of eddy-covariance flux observations for a given set of state variables. Our work reinforces this idea by providing an alternative mean estimate using a purely data-driven approach. The ANN-predicted mean is similar to existing bulk algorithms, with marginally higher statistical correlation to data. Additionally, the ANN-predicted standard deviation around the mean offers a measure of uncertainty (variability). Tests using OWS Papa data and a single-column model show that the difference in THF between the two algorithms' estimates exhibits signs of seasonality, so as the response of upper ocean states, despite of rather small magnitudes. Stochastic ensemble runs indicate that the spread of MLD is largest during the spring restratification.

We comment that the current framework can also be used to estimate the standard deviation around a given mean estimation, e.g. any existing bulk algorithm. In other words, we will deal with the residual between the observation and the bulk algorithm (or ANN) prediction. It is important to note that the uncertainty we estimate does not distinguish between measurement uncertainty (both instrument and sampling but dominantly sampling uncertainty) and missing physics (the inability to account for all predictive input variables). Such a partition is difficult and likely depends on the averaging window of the EC flux, which will be examined in future studies. Overall, the ANN-estimated uncertainty highlights regions in the input space where the data is more “scattered”. This motivates field campaigns for better sampling of these regimes and examination of mea-

surement uncertainty, as well as the use of high-fidelity numerical simulation to constrain the flux estimate (Clayson et al., 2023).

There are several possible extensions of this work in the future. The ANN model uses a limited set of input variables, which is likely the reason why the improvement over the existing algorithm is marginal. Further improvement is possible since several additional variables are measured and available in the dataset (e.g. turbulent kinetic energy), but we need to consider that many of them are not yet prognostic variables in GCMs. The height dependence is difficult to learn from the current data set alone, because there is little variance in the measurement height, but incorporating data from other measurement platforms (e.g. buoy data) might help. The probabilistic model presented assumes a uni-variate conditional Gaussian distribution for each flux component. This may not fully capture the underlying complexity, especially if the residual is induced by unrepresented physical processes. The model can be extended to a multi-variate one by including the covariance between flux components, and the Gaussian assumption can be relaxed, given that we have more data to train a more complicated statistical model.

While existing bulk algorithms represent the mean values of turbulent fluxes given limited input variables, our approach takes a step towards examining the variability around the mean values. Stochastic air-sea flux parameterization offers a promising alternative to the deterministic approach. It is crucial to prescribe the appropriate magnitude and correlation scale of noise with varying model resolutions, which needs to be better understood in future studies. In the single-column test, the spread induced by stochastic residuals exceeds that of different deterministic flux algorithms. Therefore, extending the testing to global GCMs could potentially affect the variability and address long-standing biases through interaction with horizontal transport and other nonlinear dynamics.

Open Research Section

The full NOAA PSL cruises are documented and available here <https://downloads.psl.noaa.gov/psd3/cruises/>. For this study we used a compact compilation data set https://github.com/jiarong-wu/mlflux/blob/main/fluxes_all_cruises_compilation.nc. The code for ANN training and evaluation can be found in this repository <https://github.com/jiarong-wu/mlflux>. The GOTM code is available here <https://gotm.net/portfolio/>. In particular, this work used an implementation of GOTM in the PDE solver Basilisk (Popinet, 2020) http://basilisk.fr/src/test/ows_papa.c. For comparison to bulk algorithms, we used the aerobulk-python package <https://github.com/xgcm/aerobulk-python>.

Acknowledgments

This research received support through the National Science Foundation Science and Technology Center, Learning the Earth with Artificial Intelligence and Physics, LEAP (Grant number 2019625). We thank Julia Simpson, Dr. Pierre Gentine, and Dr. Bia Villas Bôas for helpful discussions and feedback. Dr. Aakash Sane, Dr. Qing Li, and Dr. Stéphane Popinet provided valuable advice regarding the GOTM implementation. We acknowledge the continued efforts from NOAA PSL lab (including Dr. Ludovic Bariteau, Dr. Chris Fairall, and many others) in obtaining and publishing the EC datasets. This research was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

- Barnes, E. A., Barnes, R. J., & Gordillo, N. (2021). *Adding Uncertainty to Neural Network Regression Tasks in the Geosciences*. arXiv. doi: 10.48550/arXiv.2109.07250

- Berner, J., Achatz, U., Batté, L., Bengtsson, L., Câmara, A. d. l., Christensen, H. M., ... Yano, J.-I. (2017). Stochastic Parameterization: Toward a New View of Weather and Climate Models. *Bulletin of the American Meteorological Society*, 98(3), 565–588. doi: 10.1175/BAMS-D-15-00268.1
- Biri, S., Cornes, R. C., Berry, D. I., Kent, E. C., & Yelland, M. J. (2023). AirSeaFluxCode: Open-source software for calculating turbulent air-sea fluxes from meteorological parameters. *Frontiers in Marine Science*, 9. doi: 10.3389/fmars.2022.1049168
- Blay-Carreras, E., Pardyjak, E. R., Pino, D., Alexander, D. C., Lohou, F., & Lothon, M. (2014). Countergradient heat flux observations during the evening transition period. *Atmospheric Chemistry and Physics*, 14(17), 9077–9085. doi: 10.5194/acp-14-9077-2014
- Bonino, G., Iovino, D., Brodeau, L., & Masina, S. (2022). The bulk parameterizations of turbulent air-sea fluxes in NEMO4: the origin of sea surface temperature differences in a global model study. *Geoscientific Model Development*, 15(17), 6873–6889. doi: 10.5194/gmd-15-6873-2022
- Bouin, M.-N., Lebeaupin Brossier, C., Malardel, S., Voldoire, A., & Sauvage, C. (2024). The wave-age-dependent stress parameterisation (WASP) for momentum and heat turbulent fluxes at sea in SURFEX v8.1. *Geoscientific Model Development*, 17(1), 117–141. doi: 10.5194/gmd-17-117-2024
- Bourras, D., Reverdin, G., Caniaux, G., & Belamari, S. (2007). A Nonlinear Statistical Model of Turbulent Air–Sea Fluxes. *Monthly Weather Review*, 135(3), 1077–1089. doi: 10.1175/MWR3335.1
- Bradley, F., & Fairall, C. (2006). *A Guide to Making Climate Quality Meteorological and Flux Measurements at Sea* (Tech. Rep.). NOAA technical memorandum OAR PSD.
- Brunke, M. A., Fairall, C. W., Zeng, X., Eymard, L., & Curry, J. A. (2003). Which Bulk Aerodynamic Algorithms are Least Problematic in Computing Ocean Surface Turbulent Fluxes? *Journal of Climate*.
- Clayson, C. A., DeMott, C., De Szoek, S., Chang, P., Foltz, G., Krishnamurthy, R., ... Patterson, M. (2023). *A New Paradigm for Observing and Modeling of Air–Sea Interactions to Advance Earth System Prediction* (Tech. Rep.). U.S. CLIVAR Project Office. doi: <https://doi.org/10.5065/24j7-w583>
- Cronin, M. F., Gentemann, C. L., Edson, J., Ueki, I., Bourassa, M., Brown, S., ... Zhang, D. (2019). Air–Sea Fluxes With a Focus on Heat and Momentum. *Frontiers in Marine Science*, 6.
- Cummins, D. P., Guemas, V., Blein, S., Brooks, I. M., Renfrew, I. A., Elvidge, A. D., & Prytherch, J. (2024). Reducing Parametrization Errors for Polar Surface Turbulent Fluxes Using Machine Learning. *Boundary-Layer Meteorology*, 190(3), 13. doi: 10.1007/s10546-023-00852-8
- Deardorff, J. W. (1972). Numerical Investigation of Neutral and Unstable Planetary Boundary Layers. *Journal of the Atmospheric Sciences*.
- Edson, J. B., Jampana, V., Weller, R. A., Bigorre, S. P., Plueddemann, A. J., Fairall, C. W., ... Hersbach, H. (2013). On the exchange of momentum over the open ocean. *Journal of Physical Oceanography*, 43(8), 1589–1610. doi: 10.1175/JPO-D-12-0173.1
- Fairall, C. W., Bradley, E. F., Hare, J. E., Grachev, A. A., & Edson, J. B. (2003). Bulk Parameterization of Air–Sea Fluxes: Updates and Verification for the COARE Algorithm. *Journal of Climate*.
- Fairall, C. W., Bradley, E. F., Rogers, D. P., Edson, J. B., & Young, G. S. (1996). Bulk parameterization of air-sea fluxes for Tropical Ocean–Global Atmosphere Coupled-Ocean Atmosphere Response Experiment. *Journal of Geophysical Research: Oceans*, 101(C2), 3747–3764. doi: 10.1029/95JC03205
- Gleckler, P. J., & Weare, B. C. (1997). Uncertainties in Global Ocean Surface Heat Flux Climatologies Derived from Ship Observations. *Journal of Climate*,

- 10(11), 2764–2781. doi: 10.1175/1520-0442(1997)010<2764:UIGOSH>2.0.CO;2
- Guillaumin, A. P., & Zanna, L. (2021). Stochastic-Deep Learning Parameterization of Ocean Momentum Forcing. *Journal of Advances in Modeling Earth Systems*, 13(9). doi: 10.1029/2021MS002534
- Harrop, B. E., Ma, P.-L., Rasch, P. J., Neale, R. B., & Hannay, C. (2018). The Role of Convective Gustiness in Reducing Seasonal Precipitation Biases in the Tropical West Pacific. *Journal of Advances in Modeling Earth Systems*, 10(4), 961–970. doi: 10.1002/2017MS001157
- Hsu, C.-W., DeMott, C. A., Branson, M. D., Reeves Eyre, J., & Zeng, X. (2022). Ocean Surface Flux Algorithm Effects on Tropical Indo-Pacific Intraseasonal Precipitation. *Geophysical Research Letters*, 49(7), e2021GL096968. doi: 10.1029/2021GL096968
- Large, W. G., McWilliams, J. C., & Doney, S. C. (1994). Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Reviews of Geophysics*, 32(4), 363–403. doi: 10.1029/94RG01872
- Large, W. G., & Yeager, S. G. (2009). The global climatology of an interannually varying air-sea flux data set. *Climate Dynamics*, 33(2), 341–364. doi: 10.1007/s00382-008-0441-3
- Leufen, L. H., & Schädler, G. (2019). Calculating the turbulent fluxes in the atmospheric surface layer with neural networks. *Geoscientific Model Development*, 12(5), 2033–2047. doi: 10.5194/gmd-12-2033-2019
- McCandless, T., Gagne, D. J., Kosović, B., Haupt, S. E., Yang, B., Becker, C., & Schreck, J. (2022). Machine Learning for Improving Surface-Layer-Flux Estimates. *Boundary-Layer Meteorology*, 185(2), 199–228. doi: 10.1007/s10546-022-00727-4
- Nix, D., & Weigend, A. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (Vol. 1, pp. 55–60 vol.1). doi: 10.1109/ICNN.1994.374138
- Nuijens, L., Wenegrat, J., Lopez Dekker, P., Pasquero, C., O'Neill, L. W., Ardhuin, F., ... Laurindo, L. C. (2024). The Air-Sea Interaction (ASI) submesoscale: physics and impact. In *Lorentz Workshop*.
- Polichtchouk, I., & Shepherd, T. G. (2016). Zonal-mean circulation response to reduced air-sea momentum roughness. *Quarterly Journal of the Royal Meteorological Society*, 142(700), 2611–2622. doi: 10.1002/qj.2850
- Popinet, S. (2020). A vertically-Lagrangian, non-hydrostatic, multilayer model for multiscale free-surface flows. *Journal of Computational Physics*, 418, 109609.
- Schreck, J. S., Gagne, D. J., Becker, C., Chapman, W. E., Elmore, K., Fan, D., ... Wirz, C. (2024). Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications. *Artificial Intelligence for the Earth Systems*. doi: 10.1175/AIES-D-23-0093.1
- Umlauf, L., & Burchard, H. (2005). Second-order turbulence closure models for geophysical boundary layers. A review of recent work. *Continental Shelf Research*, 25(7), 795–827. doi: 10.1016/j.csr.2004.08.004
- Williams, P. D. (2012). Climatic impacts of stochastic fluctuations in air-sea fluxes. *Geophysical Research Letters*, 39(10), 2012GL051813. doi: 10.1029/2012GL051813
- Yu, L. (2019). Global Air-Sea Fluxes of Heat, Fresh Water, and Momentum: Energy Budget Closure and Unanswered Questions. *Annual Review of Marine Science*, 11 (Volume 11, 2019), 227–248. doi: 10.1146/annurev-marine-010816-060704
- Zhou, S., Shi, R., Yu, H., Zhang, X., Dai, J., Huang, X., & Xu, F. (2024). A Physical-Informed Neural Network for Improving Air-Sea Turbulent Heat Flux Parameterization. *Journal of Geophysical Research: Atmospheres*, 129(17), e2023JD040603. doi: 10.1029/2023JD040603

Supporting Information

Jiarong Wu¹, Pavel Perezhigin¹, David John Gagne², Brandon Reichl³,
Aneesh Subramanian⁴, Elizabeth Thompson⁵, and Laure Zanna¹

¹Courant Institute, New York University

²National Center for Atmospheric Research

³Geophysical Fluid Dynamics Laboratory

⁴Atmospheric and Oceanic Sciences, University of Colorado Boulder

⁵NOAA Physical Sciences Lab, Boulder, CO, USA

Contents of this file

1. Text S1 to S2
2. Figures S1 to S3
3. Tables S1

Introduction

In Text S1, we describe the training of ANNs. In Text S2, we describe the numerical experiments conducted in GOTM. In Figure S1, we compare the ANN and bulk algorithm on a 2D input plane. In Figure S2, we show the difference in upper ocean state between using KPP and $k - \epsilon$ schemes, while the fluxes are fixed. In Figure S3, we show the time

series of stochastic ensemble runs. In Table S1, we summarize the RMSE and R^2 of ANN predicted fluxes and bulk algorithm predicted fluxes.

Text S1: ANN training

We split the whole data set randomly into 80% training and 20% testing. We first train the mean network μ_θ to minimize mean square error (MSE) loss

$$L_{\text{mse}}(\theta) = \sum_{m=1}^{N_{\text{sample}}} [y_m - \mu_\theta(\mathbf{x}_m)]^2 \quad (1)$$

for maximum 10000 epochs (subject to early stopping), and then continue to train both the mean μ_θ and the variance σ_ϕ^2 networks simultaneously on negative log-likelihood loss (Equation 5 in the paper) for another maximum 10000 epochs. Although it is possible to train both the mean and the variance ANNs on negative log-likelihood directly, we found that training on the MSE loss first allows the mean ANN to capture more variability in the data.

After hyper-parameter searching, we choose to use rather small ANNs with [32,16] hidden layers as larger networks give little skill improvement. All inputs and outputs are normalized by subtracting the mean and dividing by standard deviation. Both stages of training are subject to early stopping based on the loss computed on the held-out testing dataset. The initial learning rate is 0.0005. For early stopping, the learning rate is reduced in half after 200 epochs without decrease in loss, and the training is stopped after 800 epochs without further decrease in loss.

This procedure of hyper-parameter tuning was not trivial because high-quality measurements of air-sea fluxes are sparse and the distribution of both inputs and outputs can shift significantly between subsets of data. These are challenges for data-driven models, which

we overcome through the above-mentioned model design, training strategies, and cross validation. The most important factors that promote generalizability, in our experience, is input feature selection and techniques that prevent over-fitting such as early stopping.

Text S2: the GOTM experiment

The case we run is at the location of Ocean Weather Station (OWS) Papa, which is an important long-term monitoring site in the North Pacific Ocean (145°W , 50°N). Meteorological variables and ocean state variables are measured 3-hourly, which provide inputs to the flux algorithms. Temperature and salinity profile observations are available as well. Since the uncertainty in THF is much larger than that in momentum flux, we change only the THF while using the same momentum flux as the control run. The fresh water flux, short wave and net long wave radiations are also unchanged.

We compute the THF using 3-hourly observational records and interpolate the flux to 1-hourly files. The fluxes are then read-in during run-time and linearly interpolated to integration time steps. In other words, we are running the simulations in a “forced” mode in the sense that the fluxes are de-coupled from the state variables, which is common for such setups. The temporal discretization is semi-implicit and the time stepping is chosen to be 10 minutes for both KPP and $k - \epsilon$ schemes. The vertical grid has a uniform grid spacing of 1 m for a total extension of 200 m.

As an underlying assumption for the single-column model all lateral advection terms are neglected. However, it is known that ignoring horizontal advection at the OWS Papa location results in a net heat flux imbalance on the order of 30 W/m^2 , which will cause a drift in the sea surface temperature for long simulations. Instead of finding ad-hoc compensations for the horizontal advection as in Large et al. (1994), we choose to restart

the simulation with the observed vertical profiles regularly. This is similar to the approach of Li et al. (2021), but instead of restarting four times a year, we restart every month and only examine the effects due to changing flux within the one-month windows. We have run years 2011, 2012, 2015, and 2016 where consecutive records (no gaps over 8 hours) of all input variables exist.

The correlated perturbation is generated through an auto-regressive process of order 1. According to AR(1), the perturbation ϵ at time step n is given by

$$\epsilon_n = r\epsilon_{n-1} + (1 - r^2)^{1/2}\eta_n \quad (2)$$

where $\eta_n \sim \mathcal{N}(0, \sigma_n)$. r is the lag-one auto-correlation given by $r = 1 - \Delta t/T$ where Δt is the time step and T is the auto-correlation time. In our case $\delta_t = 3$ hours and the remaining parameter to determine is the auto-correlation time T . To do this, we compute the time-lagged auto-correlation $C(\tau) = \langle Q(t)Q(t + \tau) \rangle / \langle Q^2 \rangle$ and find the e-folding time of $C(\tau)$ as the correlation time T . From the entire PSL eddy-covariance dataset T is estimated to be 60 hours, although there are likely variations between geographical locations that we are not taking into account. For each time step, the stochastic flux is then

$$\hat{Q}_n = Q_n + \epsilon_n.$$

Also note that we are imposing an additive but state-dependent stochastic perturbation, since the magnitude of ϵ depends on the predicted σ , which in turn depends on the input state variables \mathbf{X} .

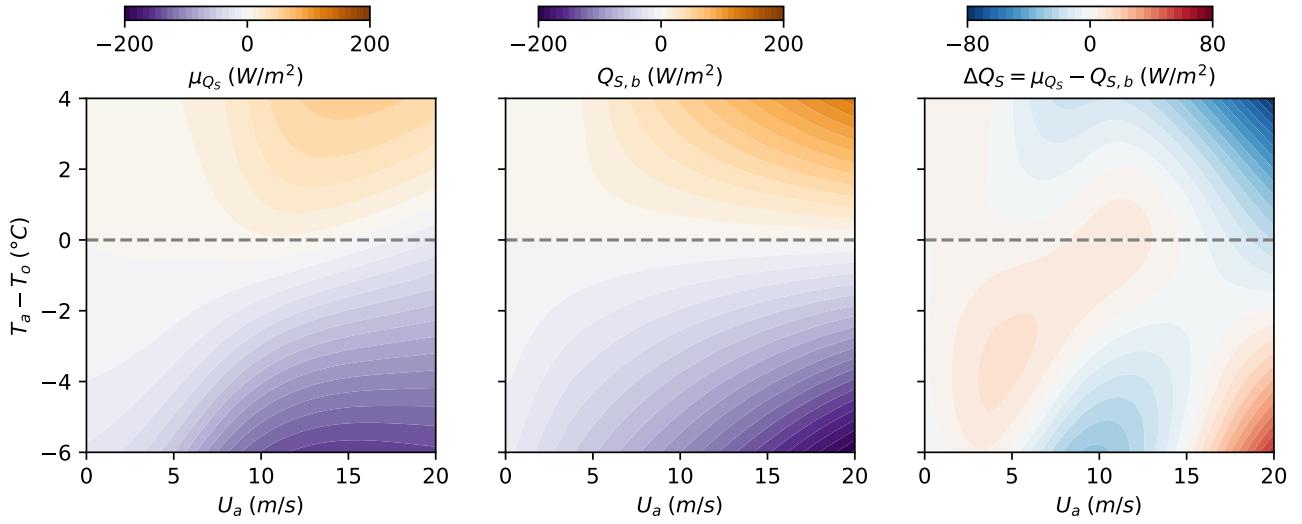


Figure S1. Structure of the prediction of sensible heat flux for bulk algorithm and for ANN, and the difference between them, for the same grid as in Figure 2(c) in the main text. The bulk algorithm is strictly down-gradient, unlike the ANN. The large differences are mainly at high wind speed and large temperature differences.

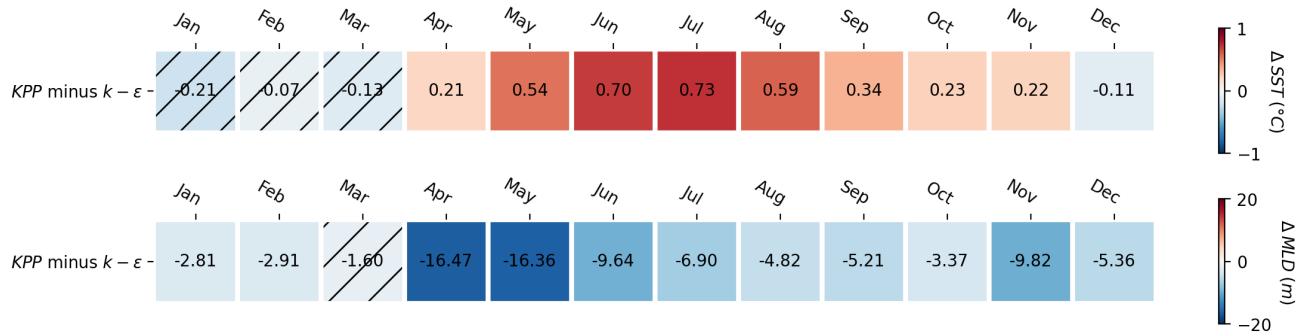


Figure S2. The bias in SST and MLD caused by changing the vertical mixing scheme from $k - \epsilon$ to KPP. KPP scheme causes a significant bias in SST and MLD, which is much larger in magnitude compared to the discrepancy caused by switching flux algorithm. The bias is also of uniform sign over the one year period, which is different from the seasonal discrepancy caused by flux algorithm.

Table S1. Statistical score of ANN prediction evaluated on subsets of the dataset, compared to COARE3.0 bulk algorithm prediction. For the cross-wind momentum flux, bulk algorithms predict zero.

		Total (10079)	Metz (3068)	N. Pac./Atl. (653)	SO (506)	Tropics (5846)
τ_x	RMSE	ANN	0.044	0.053	0.054	0.062
		Bulk	0.047	0.057	0.058	0.069
	R2	ANN	0.874	0.878	0.856	0.865
		Bulk	0.825	0.861	0.829	0.832
τ_y	Bias	ANN	0.000	-0.001	0.005	-0.001
		Bulk	-0.006	-0.005	-0.008	-0.014
	RMSE	ANN	0.034	0.039	0.030	0.052
		Bulk	0.036	0.043	0.030	0.056
Q_L	R2	ANN	0.116	0.172	0.003	0.108
		Bulk	-0.019	-0.006	0.000	-0.036
	Bias	ANN	0.000	0.001	-0.002	0.005
		Bulk	0.005	0.003	0.000	0.010
Q_S	RMSE	ANN	30.3	26.2	18.4	28.8
		Bulk	34.0	29.1	23.8	30.7
	R2	ANN	0.682	0.768	0.671	0.753
		Bulk	0.601	0.714	0.447	0.719
	Bias	ANN	0.225	0.907	-4.607	4.666
		Bulk	-5.496	-8.239	-9.465	4.538
	RMSE	ANN	15.0	17.9	9.52	16.9
		Bulk	15.7	19.2	10.9	17.1
	R2	ANN	0.577	0.671	0.830	0.786
		Bulk	0.541	0.622	0.779	0.781
	Bias	ANN	-0.073	-0.466	0.224	0.180
		Bulk	-0.361	-0.180	-0.376	1.302

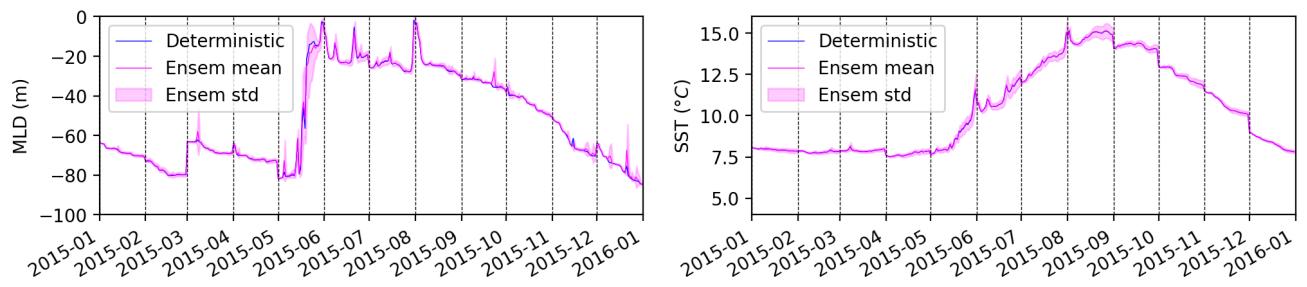


Figure S3. Time series of MLD and SST in 20 ensemble runs of year 2015. Solid line shows the ensemble mean, which is very close to the deterministic run shown in blue. The shaded envelop shows the plus minus one std, which illustrates the spread between ensemble members. The spread is shown in the heatmaps of Figure 4(d) and (e) in the main text.