Responses to the Comments and Suggestions

Reviewer 1:

Summary and Merit:

Global air-sea flux estimates are useful for understanding the transport of heat and water throughout the globe. With this dataset, the authors use a physics-constrained data-driven method to generate a dataset at moderate resolution (0.25 degrees) from 1993-2017. A key improvement is realistic representation of the ratio of SHF to LHF. While I think the work itself is a very interesting exercise and think this has strong potential to be a useful dataset, I do have a significant concern that I would like to see discussed.

Re: Thank you for your commonts. We have carefully considered all your comments and suggestions and made corresponding point-by-point responses and revisions. Specifically, reviewer comments are shown in black, our responses in blue, and the corresponding revisions in the manuscript are highlighted in red. We hope that our responses and the revised manuscript would be satisfactory.

Main comment:

I am not entirely convinced that the training dataset has large enough spatial and temporal coverage for the neural network to accurately generalize and produce a product with global-scale coverage. In particular, from Figure 2, it looks like the training observations are disproportionately from the tropical ocean. Outside of the tropics, only the northeast Pacific and North Atlantic appear to have (visually) reasonable coverage. To evaluate performance on “unseen” locations, the authors employ spatial-informed cross validation. While this procedure demonstrates that predictions are reasonably accurate at the different spatial domains that are part of the training set, this does not indicate that predictions will be accurate in regions where there are not any existing data. For instance, there are many locations in the southern hemisphere presumably characterized by different dynamics than the locations in training dataset. The comparisons between basins presented later are also only reflective of the locations in Fig 2, I think. Of additional concern is that there are many variables used in training which likely have a relationship with air-sea fluxes that is very location-specific.

I do appreciate that the authors attempt to address this issue with the above, but I don't think this goes far enough. I also acknowledge that this is not an easy comment to address (i.e., more buoy measurements cannot be used if the buoys do not exist). But, I still think the discussion of this could be improved. One idea might be to perform an even more targeted form of cross-validation, e.g., removing one of the isolated locations from training to see how well the neural network performs— and use this to quantify uncertainty. E.g., Remove the single location south of Australia from training, and see how the NN performs for predictions of that location when only the others are used in training. The current Figures 3-5 lump data together from different regions, so it is not possible to determine how well performance is for the isolated locations. Such an approach could be repeated for other single isolated locations to get a generalized idea of uncertainty at several of the remote locations not included in training. There probably could be other ways to address it as well. But in any case, there needs to be some manner of disclaimer- the R values and RMSE shown represent performance at the locations used in training and do not necessarily indicate the same performance in a generalized global sense.

Re: We appreciate the reviewer's thoughtful and constructive comments regarding the limitations in spatial coverage of the training dataset and agree that, despite our use of spatial-informed cross validation, the current approach does not fully quantify performance in truly unseen regions. Additionally, we fully agree with the reviewer's concern that the relationships between air-sea fluxes and the selected input variables may be location-specific due to regional dynamics.

In response to the reviewer's suggestion, we conducted an additional targeted cross-validation focusing on isolated locations in the high-latitude Southern Hemisphere. Specifically, we selected two buoy sites [Southern Ocean Flux buoy from the Upper Ocean Processes Group (UOP) and Global Southern Ocean Station buoy from the Ocean Observatories Initiative (OOI)], which are geographically isolated from the rest of the training dataset. We removed the data from each of these locations from the training dataset in turn and evaluated the neural network's performance at those sites. In addition, we calculated the model's statistical metrics at the two sites under spatially-informed cross-validation and made comparison with the performance under the targeted cross-validation. The resulting metrics help assess the model's extrapolation capability in underrepresented regions. In the revised manuscript, details of this analysis have been added as follows in the sixth paragraph of Section 3.5 and presented in Tables S4–S7.

"We applied a spatial 10-fold cross-validation, which provides a more generalized assessment than traditional random cross-validation, to evaluate the BrTHF model. However, it is important to acknowledge that the spatial distribution of the training dataset is inherently imbalanced, with a heavy concentration of observations in the Tropics and the Northern Hemisphere. In contrast, the Southern Hemisphere—particularly the Southern Ocean—suffers from sparse or even missing observational coverage. Given that the environmental conditions in these underrepresented or data-sparse regions may differ significantly from those captured in the training dataset, the selected input variables for the observations may lead to large uncertainty in the model's performance in these areas. To further assess the model's ability to extrapolate to such regions, we conducted an additional targeted cross-validation. Specifically, we excluded stations from the Southern Ocean [i.e., Southern Ocean Flux Station (SOFS) and Global Southern Ocean Station (GSOS)] from the training dataset and used them solely for validation. Results presented in Tables S4 and S5 show that the BrTHF model achieved the best performance in terms of LHF and β at the SOFS with lower RMSE of 15.6 W/m^2 and 0.73 and higher values of r of 0.96 and 0.34, respectively, while its SHF was slightly outperformed by ERA5 and the physics-free NN model. At the GSOS, BrTHF yielded more accurate estimates for SHF and β with RMSEs of 6.38 W/m^2 and 0.74 and values of r of 0.95 and 0.16, respectively, compared to other products, while its LHF was marginally less accurate than that of SeaFlux and the physics-free NN model. Moreover, under both spatially-informed cross-validation and targeted cross-validation, the model demonstrates comparable accuracy at the two sites, as shown in Figures S4–S7. These findings suggest that BrTHF retains competitive accuracy of SHF, LHF and β even in regions entirely excluded from training, reflecting promising generalization."

Furthermore, we now include a disclaimer in the revised manuscript emphasizing that the reported R values and RMSE reflect model performance only at locations with available observation at the end of the sixth paragraph of Section 3.5. We hope these additions address the reviewer's concerns and improve the clarity of model generalization.

"While these results are encouraging, it is important to note that the validation remains limited to a small number of sites with available observations. Therefore, the reported r values and RMSE reflect model performance in these specific locations and do not necessarily guarantee similar accuracy in broader, unobserved ocean regions."

Line-by-line comments and suggestions:

Title/abstract – It might be helpful to explicitly mention that these are bulk flux predictions

Re: Thank you for your suggestion. We have revised the title (Bowen ratio-constrained global dataset of bulk air-sea turbulent heat fluxes from 1993 to 2017) and abstract to explicitly mention that the products are bulk flux predictions.

L66 – typo seriously "imped"

Re: Thank you for your comment. We have revised “imped” to “impeded”.

L68 – change “ascribed” to “attributed”

Re: Revised as suggested.

L70-77 – I think this section should be more explicit on what the problems are with existing parameterizations

Re: Thank you for your comment. We have revised and expanded the second paragraph of Section 1 to more explicitly highlight the deficiencies in existing parameterizations. The revised text is as follows:

“More explicitly, existing parameterizations often rely on simplified assumptions about atmospheric stability and boundary layer dynamics, which may not hold under diverse environmental conditions. For instance, most bulk algorithms are optimized for moderate wind regimes, resulting in degraded performance and increased uncertainty when applied under weak wind regimes (Brunke, 2002; Jiang et al., 2024). At very high wind speeds, however, observations show that the drag coefficient can decrease due to sea spray and whitecap formation, reducing effective surface roughness and potentially biasing flux estimates (Cai et al., 2025). In addition, simplifications in the treatment of sea surface skin temperature, saturation humidity, and air density in the parameterizations can also introduce substantial uncertainty (Brodeau et al., 2017). Together, these limitations can contribute a lot to the biases in the SHF and LHF estimates and can even lead to the unphysical estimations of β, as Wang et al. (2025) reported.”

L78 – clarify what upscaling means in this context

Re: Thank you for your valuable comment. In the revised manuscript, we have clarified what upscaling means in the third paragraph of Section 1 as follows:

“Machine learning techniques have been extensively applied to upscale point-scale in-situ measurements of a single variable (such as soil moisture, roughness, or temperature) into grid-scale global datasets (Wang et al., 2023; Peng et al., 2022; O and Orth, 2021; Nelson et al., 2024; Fu et al., 2023).”

L93 – “patterns”

Re: Thank you for pointing out our typo. We have revised “pattern” to “patterns”.

L103 – I don’t understand what “their synergistic changes” refers to

Re: Thank you for your comment. We apologize for the lack of clarity in the original manuscript and have revised the sentence as follows:

“To improve the estimation of SHF, LHF, and β in a coordinative framework, we recently proposed an innovative Bowen ratio-informed data-driven model by considering the synergistic changes [on the one hand, ensuring physical consistency (i.e., SHF/LHF = β); on the other hand, achieving high-accuracy estimations of SHF, LHF, and β simultaneously] using a Random Forest (RF) technique (Wang et al., 2024).”

L107 – ambiguous whether “this work” refers to the 2024 work or the present paper

Re: Thank you for your comment. In the revised manuscript, we have specified that “this work” refers to Wang et al. (2024).

L118 – “three fold”

Re: Thank you for your comments. We have revised “three-folds” to “three fold”.

L146-161 – I think these datasets should be listed in table form, not as a long paragraph. It would make this much easier to read.

Re: Thank you for your suggestion. In the revised manuscript, we have reorganized those datasets into Table 1 to improve clarity and readability.

L202 – By forcing variables, it might be helpful to clarify that this means variables used in training the neural network

Re: Thank you for your valuable comment. By following the suggestions from you and reviewer 2, we have revised the title of Section 2.2.1 from “Forcing datasets” to “Learning datasets for training the neural network”.

L214 – not sure it’s necessary to list these out in paragraph form. To be concise it might be better to simply refer to the relevant table.

Re: Thank you for your suggestion. We would like to clarify that the information has already been summaried in the Table 1 in the original manuscript. Following your suggestion, we have removed the detailed dataset descriptions for conciseness.

L276 – I am concerned that the relationships between air sea fluxes and these 11 variables are not globally generalizable.

Re: Thank you for your comment. Please refer to our comprehensive and detailed response to your Main Comment.

L316 – Might be helpful to add a short explanation on why you chose these metrics

Re: Thank you for your suggestion. In the revised manuscript, we have added a brief explanation in the fourth paragraph of Section 2.4 as follows:

“These metrics—BIAS, RMSE, and r—comprehensively evaluate model performance, representing systematic deviation, dispersion between observations and estimates, and the strength and direction of the linear relationship, respectively.”

L363-383, Fig 5 – While performance in terms of RMSE is clearly improved as explained, depending on the application it might be considered a deficiency that BrTHF does not reproduce extreme values of Bowen ratio that we know exist from the observations (i.e. the distribution is not necessarily better represented than the other models). I think this needs to be explicitly discussed.

Re: Thank you for your comment. We acknowledge that our model does not fully capture the extreme values of β, which is a deficiency to be addressed in future work. However, from Figure 5, we would like to clarify that, although our model predicts β within ±2—slightly narrower than the observed range of ±5, other models and products, while capable of reaching ±5, generate numerous β values far beyond the observed range (e.g., 5 to 500 or –5 to –500). The distribution of β predicted by the BrTHF model is overall relatively better aligned with the observations compared to other products and models.

In short, although the BrTHF model slightly underestimates the extreme values of β, it avoids the occurrence of unrealistic outliers seen in other products, making it overall better aligned with observations.

In the revised manuscript, we have now explicitly discussed this limitation of β in the eighth paragraph of Section 3.5 as follows:

"While incorporating the constraint of β into the model effectively suppresses outliers, it also compresses the physically plausible range of β. As a result, the distribution of β shown in Figure 5(i) differs notably from other products and models, which may limit the product's applicability for users interested in extreme β values. It is highlighted that although the BrTHF model slightly underestimates the extreme values of β, it avoids the occurrence of unrealistic outliers (e.g., 5 to 500 or –5 to –500) seen in other products, making it overall better aligned with observations. Moving forward, we aim to enhance the model's ability to preserve physically plausible extremes while maintaining robustness against outliers in future updates.

L400+ - I think it might be useful to compare the performance by basin to the amount of data coverage between basins. This might help explain why the model performed the way it did.

Re: Thank you for your suggestion. As recommended, we evaluated several indicators of the data coverage across ocean basins, including number of buoys, number of samples, buoy density, sample desity, nearest neighbor distance (NND, the distance between a given point and its closest neighboring point) and standard deviation of NND in Table S6. By computing NND for all sample points and then calculating the mean and standard deviation, we can characterize the density and spatial uniformity of the point distribution. In general, a higher mean indicates a sparser distribution, whereas a higher standard deviation reflects greater spatial heterogeneity.

These indicators were then used to represent data coverage across basins and, in combination, to compare model performance among different ocean basins. In the revised manuscript, the relevant findings have been incorporated into the fifth paragraph of Section 3.5 as follows:

"Based on Figure 2 and Table S6, we observe that the spatial coverage of observations varies across different ocean regions: the Northern Hemisphere generally has higher coverage than the Southern Hemisphere, with the Northern Pacific Ocean exhibiting the highest coverage, while the Arctic Ocean shows the lowest. Comparing spatial coverage with accuracy metrics reveals a more complex relationship between model performance and data coverage. Specifically, the values of r tend to be lower in regions with lower coverage — a pattern consistent across SHF, LHF, and β. However, RMSE does not follow this trend. For SHF and β, RMSEs in the Northern Hemisphere are generally higher than those in the Southern Hemisphere. Similarly, for LHF, RMSEs are higher in the Northern Hemisphere except in the Indian Ocean, where the pattern differs."

Fig 7 – I would recommend to use a color other than blue for the second and third columns. As is, it is confusing that dark blue = poor performance in column 1, but dark blue = good performance in columns 2 and 3.

I also think it should be very clear that the basins here just represent the buoy locations that are available in those basins; not uniform coverage in them.

Re: Thank you for your suggestion. In the revised manuscript, we updated the color schemes in the second and third columns to a diverging colormap for more consistent interpretation. We also clarified in the caption of the Figure 7 that the displayed ocean basins only reflect the locations of available buoy observations rather than uniform coverage as follows:

"It should be noted that the statistical metrics for each ocean basin were calculated using observations from the available buoys within the corresponding basin."

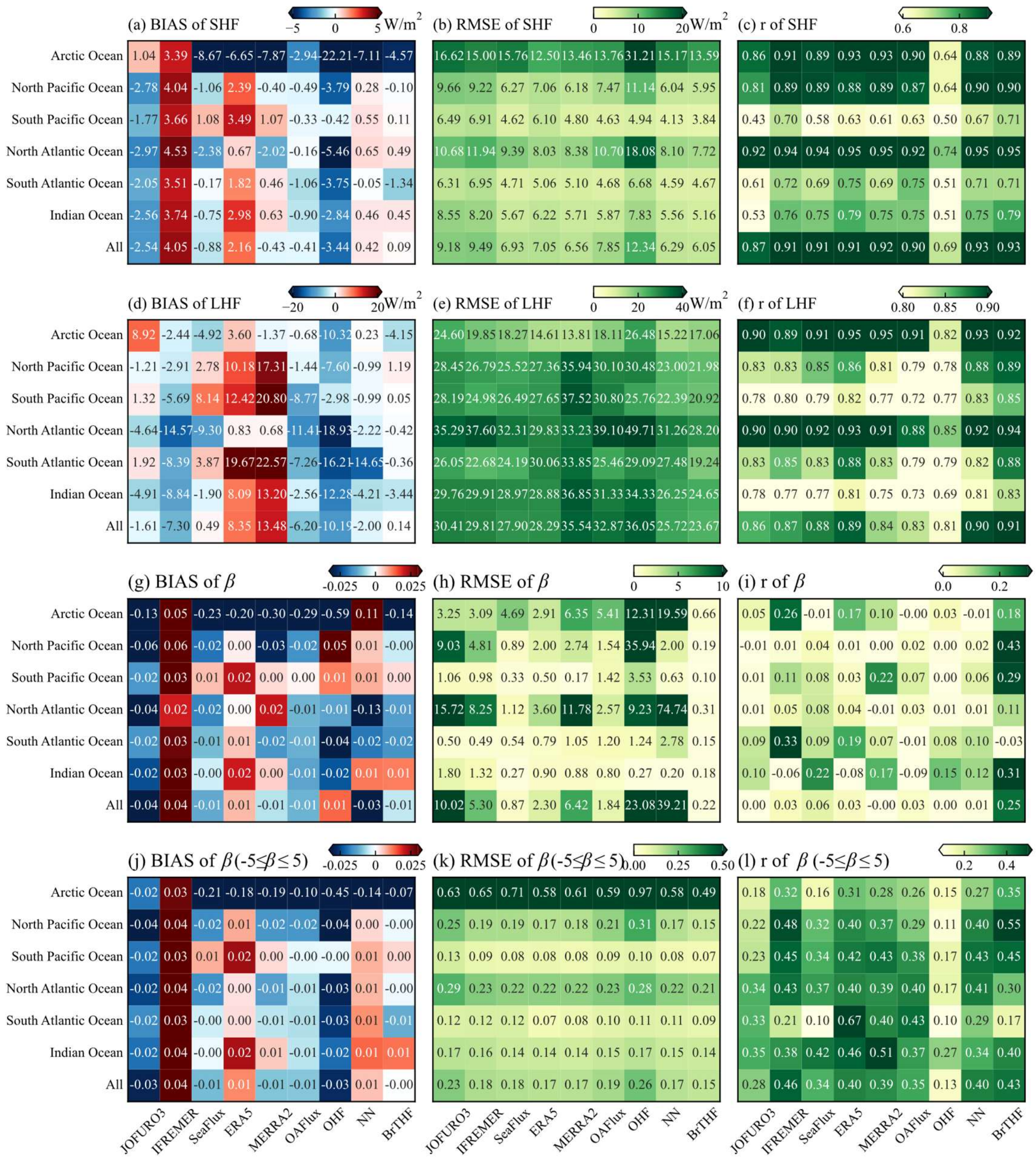

Figure 7. Heatmaps of BIAS, RMSE and r metrics for the validation of estimated daily SHF (a - c), LHF (b - e), β (f - i) and β ($-5 \leq \beta \leq 5$, j - l) from the BrTHF model, the physics-free NN models and the seven products against the in-situ observations across different ocean basins. It should be noted that the statistical metrics for each ocean basin were calculated using observations from the available buoys within the corresponding basin

L448-449 – That looks true for all datasets, not just BrTHF from Figure 8. I would recommend to clarify.

Re: Thank you for your comment. We agree that the less pronounced peak in SHF and β compared to LHF is observed across all products in Figure 8, not just BrTHF. The sentence has been revised to clarify this seasonal pattern.

Fig 8-9 – Is there a measure of uncertainty in these long-term averages that could be included on the plots?

Re: Thank you for your suggestion. We chose the commonly used standard deviation to represent uncertainty of the long-term averages and have added it to Figures 8 and 9 as follows:

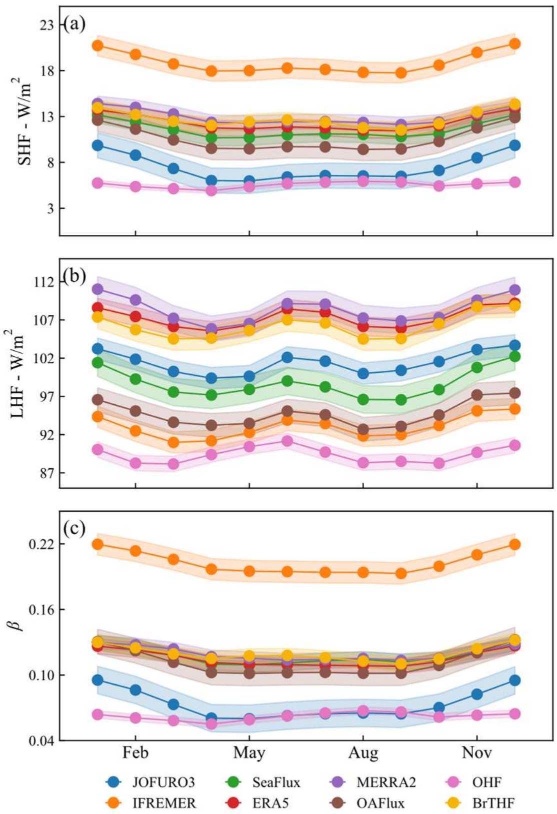

Figure 8. Intra-annual cycles of area-weighted global monthly mean of SHF (a), LHF (b) and *β* (c) from the eight products from 1993 to 2017. The shaded areas indicate ±1 standard deviation around the mean.

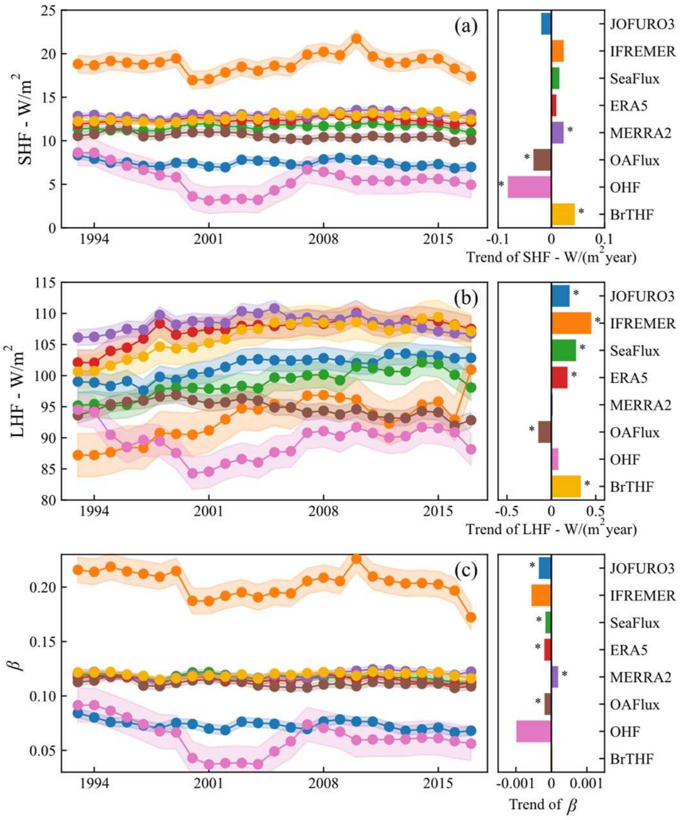

Figure 9. Inter-annual evolution of area-weighted global mean SHF (a - b), LHF (c - d) and β (e - f) from 1993 to 2017. The trends were calculated based on the Sen's slope method. The * in the sub-figures (b, d and f) represent the trend passed the Mann-Kendall significant test ($p < 0.05$). The shaded areas indicate ±1 standard deviation around the mean.

L472 – "rest of the products"

Re: Thank you for your suggestion. We have revised "the rest five products" to "rest of the products".

L482-483 – I would recommend to speculate on what regions/mechanism may have caused this positive trend, as it differs from the other products.

Re: Thank you for your comment. As shown in Figure 9, the differences between trends in SHF and LHF from BrTHF product were relatively lower than those from other products. In contrast, except for MERRA2, other products show a stronger increasing trend in LHF than in SHF (e.g., IFREMER, SeaFlux, and ERA5), or an increasing trend in LHF accompanied by a decreasing trend in SHF (e.g., JOFURO3, OAFlux, and OHF). This is likely the cause of the different β trend in BrTHF (weakly positive, close to zero, and not statistically significant), and such differences can be further attributed to disparities in the accuracy of SHF, LHF, and β among the products. Considering that our validation results indicate higher overall accuracy of BrTHF product, the β trend in our product may be reasonable. Nevertheless, the reliability of long-term trends ultimately requires further observational data to determine which product provides the most accurate representation.

In the revised manuscript, we have clarified the possible reason in the third paragraph of Section 3.2 as follows:

"However, the BrTHF product exhibited a weak positive trend, which may be attributed to the relatively smaller differences between the SHF and LHF trends in BrTHF compared to those in other products."

Sec 3.3 – This section implies that performance between BrTHF and Seaflux-ERA5 is similar, even in regard to Bowen ratio which earlier seemed to be the point of significant improvement for BrTHF. Please comment on this.

Re: Thank you for your comment. We would like to clarify that the large-scale spatial patterns of air-sea turbulent heat fluxes are primarily shaped by atmospheric circulation and sea surface properties (e.g., sea surface temperature, and salinity), which result in broadly similar spatial structures across different products as the reviewer pointed out. However, notable differences remain as shown in the difference maps (first and second rows, fourth column) and scatter plots (fourth row, first and second columns) of Figures 10-12. For instance, BrTHF shows significantly higher SHF values in the high-latitude Northern Hemisphere compared to SeaFlux, with greater dispersion in the scatter plots. These spatial and statistical differences reflect the improvements achieved by our model and have been described in Section 3.3 of the original manuscript.

In the revised manuscript, we have added a discussion in Section 3.3, third paragraph, to clarify the potential explanation as follows:

"In addition, the OHF product did not reproduce similar large-scale spatial patterns of air–sea turbulent heat fluxes observed in BrTHF, ERA5, and SeaFlux, which are primarily shaped by atmospheric circulation and sea surface properties (e.g., sea surface temperature and salinity)."

Fig 13 – It's a bit confusing that the labels on the color bar are below the plots on the left. It might be more intuitive to add a title above each subplot rather than a colorbar label.

Re: Thank you for your suggestion. In the revised manuscript, we have moved the labels to the top-left corner of each subplot in Figure 13 to improve readability and make the figure more intuitive.

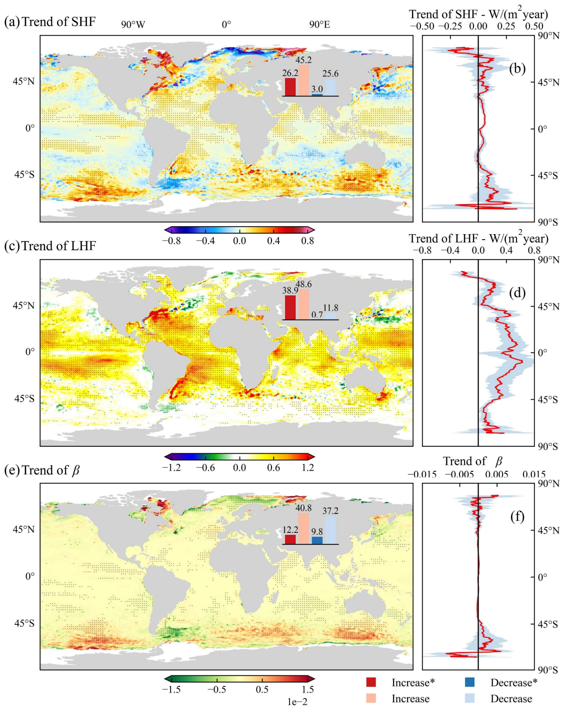

Figure 13. Spatial maps of inter-annual trends for SHF (a), LHF (c), and β (e) from the BrTHF product for the period 1993 to 2017. The trends were calculated using the Sen's slope method. Dotted areas indicate oceans where the p-value of the Mann-Kendall significance test is less

than 0.05. Panels (b), (d) and (f) represent the inter-annual trends of zonal annual averages for SHF, LHF and β, respectively.

L553-555 – Do we trust these results, considering that there was significant uncertainty at high latitudes (and the NN was trained on few observations from high latitudes)? Could this be an artifact of the training data/procedure?

Re: Thank you for your comment. Due to the scarcity of long-term observations in high-latitude oceans, we assessed the reliability of simulated trends of BrTHF in these regions by comparing them with the corresponding trends from seven widely used products. As shown in Figures S2–S4, in these high-latitude regions, the trends simulated by the BrTHF are largely consistent with those of most other products—for example, SHF exhibits a pronounced increase in the Kara Sea, Gulf Stream, Baffin Bay, Brazil Current, Sea of Okhotsk, and Sea of Japan, with differences mainly in magnitude. Given that these products are developed based on physically well-founded models, we consider the trends simulated by our product to be reliable.

In the revised manuscript, we have added a discussion about the reliability of simulated trends in the fourth paragraph of Section 3.5 as follows:

"The generalization capability of the model can also affect the accuracy of simulated long-term trends. In Figure 13, we present the spatial distributions of long-term trends for SHF, LHF, and β simulated by the BrTHF product. Considering the scarcity of training data in high-latitude oceans, the simulated long-term trends in these regions may be associated with larger uncertainties. However, due to the lack of long-term observations in high-latitude oceans, we cannot validate the simulated trends using observational records as done in previous studies for mid- and low-latitude regions (Tang et al., 2024; Weller et al., 2022). To address this, we examined the spatial distribution of long-term trends from the other seven widely used products. Specifically, in these high-latitude regions, the trends simulated by the BrTHF are largely consistent with those of most other products—for example, SHF exhibits a pronounced increase in the Kara Sea, Gulf Stream, Baffin Bay, Brazil Current, Sea of Okhotsk, and Sea of Japan, with differences mainly in magnitude."

L588 – "custom"

Re: Thank you for your comment. We have revised "customed" to "custom".

L590 – I'm unconvinced that the absence of outliers is an improvement, since outliers exist in the observations. Please comment on this.

Re: Thank you for your comment. We acknowledge that outliers do exist in observations; however, many of the outliers are likely caused by measurement errors. Considering that such outliers can severely impede model training and evaluation, we deemed it necessary to constrain the β in a reasonable range to enable simultaneous high-accuracy estimation of SHF, LHF, and β.

Specifically, we calculated the cumulative distribution of daily β for each product and their ensemble (across all products). The medians of the 1^{st} and 99^{th} percentiles, approximately -5 and 5, respectively, were selected as the minimum and maximum of valid daily β, as shown in Figure S1. We did not derive the constraints of β directly from observations, primarily because the limited spatial coverage of observations may not provide a range that is generally applicable across all ocean basins. While simulated data offer global representativeness, they may also contain outliers. Therefore, we manually set a reasonable β range based on the 1^{st} -99^{th} percentiles (in ascending order), as already presented in the fifth paragraph of Section 2.1. This range provides a robust basis for model development, ensuring that SHF, LHF, and β can be jointly estimated with high accuracy.

In the revised manuscript, we have clarified the importance of absence of β outliers in the fifth paragraph of Section 2.1 as follows:

"Although outliers exist in observations, some are likely caused by measurement errors. Considering that such outliers can severely impede model training and evaluation, it was necessary to constrain β within a reasonable range to enable simultaneous high-accuracy estimation of SHF, LHF, and β."

L609-618 – I'm not sure that this isn't also true for the present dataset based on looking at Figure 2

Re: Thank you for your comment. This issue appears closely related to model generalization and has been discussed in detail in the Main Comment.

L666 – Performance in terms of SHF/LHF did not clearly look superior based on the plots. Please clarify that the largest improvement is in Bowen ratio.

Re: Thank you for your comment. In the revised manuscript, we have clarified that the most significant improvement achieved by the BrTHF model is in the estimation of the β, while its performance in estimating SHF and LHF is generally comparable to or slightly better than other models and products in the second paragraph of Section 5 as follows:

"The BrTHF model demonstrated the most significant improvement in estimating the β, while its performance in estimating SHF and LHF was generally comparable to or slightly better than that of the physics-free NN models and the seven widely used air-sea turbulent heat products (including the JOFURO3, IFREMER, SeaFlux, ERA5, MERRA2, OAFlux and OHF products)."

Reference：

Brodeau, L., Barnier, B., Gulev, S.K. and Woods, C., 2017. Climatologically Significant Effects of Some Approximations in the Bulk Parameterizations of Turbulent Air–Sea Fluxes. Journal of Physical Oceanography, 47(1): 5-28.

Brunke, M.A., 2002. Uncertainties in sea surface turbulent flux algorithms and data sets. Journal of Geophysical Research, 107(C10).

Cai, L., Wang, B., Wang, W. and Feng, X., 2025. The Impact of Air–Sea Flux Parameterization Methods on Simulating Storm Surges and Ocean Surface Currents. Journal of Marine Science and Engineering, 13(3).

Jiang, Y., Li, Y., Lu, Y., Wu, T. and Gao, Z., 2024. Evaluating modifications to air–sea momentum flux parameterizations under light wind conditions in CAM6. Climate Dynamics, 62(10): 9687-9701.

Tang, R., Wang, Y., Jiang, Y., Liu, M., Peng, Z., Hu, Y., Huang, L. and Li, Z.-L., 2024. A review of global products of air-sea turbulent heat flux: accuracy, mean, variability, and trend. Earth-Science Reviews, 249.

Wang, Y., Tang, R., Liu, M., Huang, L. and Li, Z.-L., 2025. Bowen ratio-constrained global dataset of air-sea turbulent heat fluxes from 1993 to 2017. Earth System Science Data Discussions, 2025: 1-41.

Weller, R.A., Lukas, R., Potemra, J., Plueddemann, A.J., Fairall, C. and Bigorre, S., 2022. Ocean Reference Stations: Long-Term, Open-Ocean Observations of Surface Meteorology and Air–Sea Fluxes Are Essential Benchmarks. Bulletin of the American Meteorological Society, 103(8): E1968-E1990.