Summary of major revisions

We sincerely appreciate this opportunity to revise our manuscript and we are truly grateful to the Editor and the two reviewers for their valuable comments and thoughtful suggestions on our manuscript (ID: essd-2025-272). We have carefully considered all these comments and made point-by-point responses.

In response, we added a disclaimer in Section 3.5.2 cautioning users about interpreting the BrTHF results in data-sparse high-latitude regions and discussed the application scope of β and the rationale for constraining the extreme values of β to ensure model stability and accuracy in Sections 3.5.1 and 3.5.4, respectively.

We also added analyses of model sensitivity and uncertainty, including the impact of Bowen ratio constraints (Table S9), sensitivity to input parameters (Table S10), differences between daily- and high-frequency-derived flux observations in model evaluation (Table S11), and effects of parameterization differences on model accuracy (Table S12), all presented in Section 3.5.3. In addition, we corrected errors and clarified ambiguous statements throughout the manuscript.

To improve readability, the original discussion (Section 3.5 in the previous manuscript) was reorganized into four subsections (3.5.1 Advantages, 3.5.2 Generalizability, 3.5.3 Sensitivity and Uncertainty, and 3.5.4 Limitations and Recommendations), and content on the OHF product was moved to Section 3.1.2.

The responses are marked in blue color and the revisions in red color, and we hope our revised manuscript could satisfy the reviewers.

Responses to the Comments and Suggestions

Editor

Public justification (visible to the public if the article is accepted and published):

The authors have submitted a revision that partially addressed the concerns highlighted by the reviewers. However, there are still some concerns remaining with respect to the approach.

Specifically:

- sensitivity of the model to input parameters and the effect of the models apparent limitation in predicted Bowen ratios > an exploration of the sensitivity here is merited
- the model's ability to capture dynamics in high latitude regions, which are potentially driven by ocean and atmospheric dynamics that are different from dynamics for the buoys in the training dataset > this should be discussed more clearly.
- a concern that daily aggregation of fluxes in the comparison introduces a bias in the presented results.

These and other comments from reviewer 1 should be addressed before publication.

Re: We thank the Editor for the careful assessment of the revised manuscript and for clearly summarizing the remaining key concerns. In response, we have conducted additional analyses and revised the manuscript to address all three points raised, including the sensitivity of BrTHF to input parameters and Bowen ratio limitation, the model's applicability and limitations in high-latitude regions, and the potential bias introduced by daily temporal aggregation in comparisons.

In addition, all comments and suggestions from reviewer 1 have been carefully considered and addressed in the revised manuscript, with corresponding revisions clearly indicated. We believe that these revisions improve the clarity and robustness of the study and adequately address the remaining concerns regarding the proposed approach.

Reviewer #1

Suggestions for revision or reasons for rejection

(visible to the public if the article is accepted and published)

I appreciate the significant effort put forth by the authors in revision which has improved the manuscript. However, I still have several concerns. Please see my comments below.

1) My first comment is in regard to whether the product can be trusted in regions far away from buoy observations. I appreciate that the authors have adopted my suggestion to perform targeted cross-validation of an isolated buoy location, and determine that the performance is similar to other products for remote locations. This addition is most welcome, and demonstrates that the NN exhibits some degree of generalizability that was not shown in the original version. While this is a step in the right direction, I think (and the authors acknowledge) that spatial limitations of the training data still likely influence the results. Because of this, I question whether BrTHF is an improvement over existing products on a global scale. At minimum I think a stronger disclaimer is needed considering that the authors present this as a global product.

Re: Thank you for your comment. We agree that the spatial limitations of the training data remain an important factor influencing model performance, particularly in regions far from buoy observations. While the targeted cross-validation experiments provide encouraging evidence that BrTHF can generalize to selected remote locations, we fully acknowledge that this does not constitute comprehensive validation across all poorly observed ocean regions. To address this concern, we have further strengthened the disclaimer in the second paragraph of Section 3.5.2 in the revised manuscript as follows: “Consequently, the BrTHF product should be viewed as being primarily optimized within the geographical coverage of existing buoy networks. In remote regions far from the observation-rich regions, such as the high-latitude Southern Ocean, the lack of direct ground-truth constraints may result in certain uncertainties. Users should therefore exercise caution when interpreting the global-scale performance, particularly in data-sparse basins where spatial sampling limitations are most pronounced.”

2) I think there needs to be a dedicated discussion in the text on why it is important to represent the Bowen ratio accurately. That is, what specific applications would this be useful for? For example, are there deficiencies in previous studies utilizing other products that would be resolved if the Bowen ratio was more accurately modeled (e.g., without the extreme outliers that exist in alternate products)? In terms of demonstrating that this is a useful dataset, I think this is essential to discuss. From the tables and as noted by the other reviewer, the quantitative improvement of individual SHF and LHF terms is incremental compared to other products and probably not useful on its own. The main "improvement" is in the Bowen ratio, yet I'm not sure I understand exactly why this improvement would be useful.

Re: Thank you for your comment. We agree that the importance of accurately representing β and its applications merits further discussion. The key applications and significance of β are summarized as follows:

1) Ensuring physical consistency in energy partitioning. According to the surface energy balance framework, β is the fundamental indicator of how available energy (Rn − G) is partitioned between SHF and LHF. The improvement in β can therefore lead to more physically sound and accurate estimates of SHF and LHF (Yang et al., 2025).
2) Serving as a "physical fingerprint" of climate variability. The β anomaly can act as a diagnostic tool for large-scale climate modes. For instance, β has been shown to be highly correlated with the ENSO index with a predictable 12-month lag (Jo, 2002). Accurate representation of β can thus capture predictive signals of climate variability that may be obscured when considering individual flux components alone.
3) Constraining global precipitation sensitivity. The intensification of the global hydrological cycle is constrained by the surface energy budget, and the sensitivity of precipitation to warming depends critically on β. Accurately representing β is therefore essential for reducing uncertainties in projecting the hydrological response to global warming (Wang et al., 2021).

Corresponding clarifications have been added to the first paragraph of Section 3.5.1 of the revised manuscript.

"The primary advantage of the BrTHF model lies in its accurate estimation of β, which shows the most pronounced improvement among all flux components. As a key indicator of surface energy partitioning, β is widely used within the surface energy balance framework to ensure physically consistent and reliable estimates of SHF and LHF (Yang et al., 2025). In addition, β serves as an effective diagnostic variable in the studies of large-scale climate variability (e.g., ENSO) (Jo, 2002) and in investigations of how surface energy constraints regulate the hydrological cycle (e.g., precipitation) (Wang et al., 2021). With its enhanced representation of β, the BrTHF product is expected to provide more reliable support for such applications."

3) There are still some features of the predicted values that need further explanation. I agree that extreme values of Bowen ratio are eliminated in BrTHF (though this needs some more justification on why those are a problem- the authors state they result from measurement error, but I don't understand how that can be the case for model-derived products). My concern is the lack of representation of the dynamic range of Bowen ratios. The authors state that this is a slight underestimate, while from Fig 5 it seems like a large underestimate. Related, Fig 6 seems to show that for small Bowen ratios, the other products yield a more realistic range of values than BrTHF (if I'm interpreting that figure correctly). So essentially, it seems like BrTHF is eliminating one problem (extreme outliers) at the expense of creating another (too small of a dynamic range). The second problem seems to also be by intention (L213)- i.e., intentionally training on a narrower range – I don't understand how this is reasonable to do. These aspects need to be clearly discussed, and there needs to be an explanation on whether this trade-off is actually an improvement for potential future applications using the dataset.

Re: Thank you for your comment. We agree that the treatment of extreme values of β and the resulting reduction in dynamic range require clearer justification, and that the implications of this trade-off for potential applications need to be explicitly discussed. Below we clarify our rationale.

Regarding the origin of extreme values of β, we acknowledge that they may arise from two different sources: (1) physically plausible but rare extreme conditions in the real world, and (2) numerically unstable or poorly constrained estimates associated with uncertainties in model-derived fluxes and input variables. In particular, because β is defined as the ratio of SHF to LHF, small absolute errors in either flux, especially when LHF approaches zero, can be strongly amplified, leading to spuriously large or small β even when the underlying flux estimates are reasonable. In practice, it is therefore difficult to reliably distinguish physically meaningful extremes from numerically unstable outliers.

As shown in Figure S1, such extreme values of β occupy only a very small fraction of the global distribution. To ensure robust model training and stable performance under the vast majority of conditions (approximately the 1-99% of the β distribution), we restricted the training range using a conservative β constraint (-5 to 5). This choice inevitably compresses the standard deviation and dynamic range of the predicted β, but it substantially improves model stability and accuracy for the dominant range of global ocean conditions, as shown in Figures 5 and 6.

We have now explicitly discussed this trade-off in the revised manuscript and clarified that this choice represents a balance between suppressing unrealistic outliers and preserving meaningful variability. Importantly, we view this strategy as an interim solution rather than a final one. With future improvements in the quality and spatiotemporal representativeness of observational datasets, the physically plausible extreme tails of the β distribution could be better constrained and incorporated into model training, allowing the dynamic range of β to be expanded. Corresponding caveats and future expectations have been added to the first paragraph of Section 3.5.4 in the revised manuscript.

“First, extreme values of β may arise either from physically plausible but rare ocean conditions or from numerical instability or poorly constrained estimates associated with uncertainties in model-derived fluxes and input variables. In practice, these two sources are difficult to distinguish. To ensure robust model training and stable performance across the vast majority of β conditions (approximately the 1-99% of the β distribution), we applied a conservative β constraint (-5 to 5) during training. This compresses the standard deviation and constrains the extreme tails of the predicted β, but substantially improves model stability and accuracy for the majority of ocean conditions. Although this choice leads to a narrower dynamic range compared to some other products, it ensures that the 1-99% of the β distribution is well-represented for most practical applications. It should be noted that this strategy represents an interim solution rather than a final one. With future improvements in the quality and spatiotemporal representativeness of observational datasets, physically plausible extremes (e.g., 0-1% and 99-100% of the β distribution) could be better constrained and incorporated into model training, allowing expansion of the dynamic range of β."

4) I still notice some errors and unclear statements in the text and think this would benefit from another close read-through. A few examples:

Re: Thank you for your comment. We have carefully revised the specific issues and conducted an additional thorough read-through of the entire manuscript to improve clarity, consistency, and wording. All identified errors and unclear statements have been corrected accordingly.

L72 – I don't understand what the "key process" is?

Re: Thank you for your comment. We have removed the term "key process" to avoid potential confusion, without affecting the original meaning or the clarity of the manuscript.

L106 – These aren't really separate "approaches", just studies focusing on different variables.

Re: Thank you for your comment. We have revised the "approaches" to "cases".

L211 – "measurement errors" is very vague. Also not sure it is reasonable to assume that since you are using published data, unless there's specific information in the metadata on this

Re: Thank you for your comment. We have revised the sentence in the fourth paragraph of Section 2.1 as follows:

"Note that outliers of β present in the observations are likely associated with uncertainties in the model-derived estimates and input data."

L365 – Add additional justification on the metrics. Particularly, all of these will be sensitive to extreme values, and it doesn't necessarily seem fair to remove extremes from training then evaluate in this way without some more justification.

Re: Thank you for your comment. We have added justification for the metrics in the fourth paragraph of Section 2.4 in the revised manuscript as follows:

"Note that RMSE and r can be sensitive to extreme values, particularly for β, which is a ratio-based variable and may exhibit unrealistically large magnitudes under near-zero flux conditions. To ensure a fair evaluation, model performance is assessed both with all samples retained and with extreme β values excluded in subsequent analyses. This dual evaluation allows us to quantify overall model performance while explicitly accounting for the influence of rare extreme cases."

L605 – The Gulf Stream, Brazil Current, and Sea of Japan are not high latitude

Re: Thank you for your comment. We have revised the text to use the term 'mid- and high-latitude oceans' to accurately encompass these specific areas.

Reviewer #2

Suggestions for revision or reasons for rejection

(visible to the public if the article is accepted and published)

In accordance with journal policies, I am acting as a second reviewer on this manuscript. My review is also informed by an email sent to me from someone who was invited to serve as a reviewer, but did not have sufficient time to conduct a formal review (and thus sent me their comments in email form).

- I think that the question about representativeness/ applicability of the model in high latitudes is a major issue and merits further discussion.

Re: Thank you for your comment. Similar points were also raised by Reviewer 1, and we have provided corresponding responses. Specifically, regarding representativeness, please refer to our response to Comment 1 from Reviewer 1, and the related modifications are reflected in the second paragraph of Section 3.5.2 in the revised manuscript.

"Consequently, the BrTHF product should be viewed as being primarily optimized within the geographical coverage of existing buoy networks. In remote regions far from the observation-rich regions, such as the high-latitude Southern Ocean, the lack of direct ground-truth constraints may result in certain uncertainties. Users should therefore exercise caution when interpreting the global-scale performance, particularly in data-sparse basins where spatial sampling limitations are most pronounced."

Regarding the discussion of applications, we have addressed this in our response to Comment 2 from Reviewer 1; the corresponding revisions are included in the first paragraph of Section 3.5.1 in the revised manuscript.

"The primary advantage of the BrTHF model lies in its accurate estimation of β, which shows the most pronounced improvement among all flux components. As a key indicator of surface energy partitioning, β is widely used within the surface energy balance framework to ensure physically consistent and reliable representations of SHF and LHF estimates (Yang et al., 2025). In addition, β serves as an effective diagnostic variable in studies of large-scale climate variability (e.g., ENSO) (Jo, 2002) and in investigations of how surface energy constraints regulate the hydrological cycle (e.g., precipitation) (Wang et al., 2021). With its improved estimation of β, the BrTHF product is expected to provide more reliable support for such applications."

- I also think that the effect of the Bowen-ratio limitation on fluxes is a major question with respect to sensitivity to results.

Re: Thank you for your comments. To evaluate the sensitivity of SHF and LHF estimates to the Bowen-ratio limitation, we conducted a series of sensitivity experiments in which the allowable range of β was progressively relaxed. Specifically, the constraint of β was relaxed from a highly restrictive interval of [−1, 1] to a highly permissive range of [−50, 50], with intermediate ranges of [−5, 5] (adopted in the previous manuscript), [−10, 10], and [−20, 20], while all other model configurations were kept unchanged.

The results demonstrate that the BrTHF model-derived SHF and LHF exhibit weak sensitivity to the Bowen-ratio limitation. Specifically, the maximum variation of RMSE is approximately 0.17 W/m^2 for SHF and 0.7 W/m^2 for LHF, corresponding to less than 3% of their respective mean values of RMSE. Meanwhile, the values of r remain unchanged for both SHF (0.9) and LHF (0.91). Moreover, the values of BIAS for both SHF and LHF fluctuate around zero without systematic dependence on the imposed β range.

We have incorporated these results into the first paragraph of Section 3.5.3 in the revised manuscript, with the corresponding metrics summarized in Table S9.

"Given the key role of the β constraint in the BrTHF model, it is also important to assess the sensitivity of the estimated SHF and LHF to the imposed β range. A series of sensitivity experiments with progressively relaxed β constraints indicate that the BrTHF model exhibits weak sensitivity to the specific choice of the β range (see Table S9). The resulting variations in RMSE are small relative to their mean values, while values of r and BIAS remain largely unchanged across different β configurations. These results suggest that the improved performance of BrTHF is not driven by a particular predefined range of β, but instead reflects the robustness of the Bowen ratio–constrained machine learning framework."

Table S9. Sensitivity of BrTHF-derived SHF and LHF performance metrics to different ranges of β constraint.

β	SHF			LHF		
	BIAS (W/m^2)	RMSE (W/m^2)	r	BIAS (W/m^2)	RMSE (W/m^2)	r
[-1,1]	-0.12	6.1	0.9	-0.36	23.97	0.91
[-5,5]	0.09	6.05	0.9	0.14	23.67	0.91
[-10,10]	-0.17	6.11	0.9	-0.15	23.78	0.91
[-20,20]	-0.14	6.16	0.9	-0.76	24.27	0.91

[-50,50]	0.07	6.13	0.9	0.24	24.4	0.91

- In general, there is an issue about the sensitivity of the model to its input parameters (other than delta_T and delta_q (effectively a bulk formulation). To increase trust in the model output, some sensitivity analysis should be presented.

Re: Thank you for your comments. To assess the sensitivity of the BrTHF model to its input parameters including $diff_Q$, *and* $diff_T$, we conducted a set of one-at-a-time sensitivity experiments. Specifically, each environmental input (*SLP*, *LW*, *SW*, *SSS*, *ADT*, *CS*, *WS*, *SWH*, T_p, $diff_Q$, and $diff_T$) was independently perturbed by ±5%, ±10%, and ±20%, while all other inputs were kept unchanged.

The results (Table S10) indicate that the BrTHF model is generally insensitive to perturbations in most input parameters, with noticeable sensitivity primarily associated with *WS*, $diff_Q$, and $diff_T$. For SHF, *WS* and $diff_T$ show relatively larger influences: under moderate perturbations (±10%), RMSE increases from 6.05 to more than 6.17 (+1.98%) and 6.14 (+1.49%), respectively, while under extreme perturbations (±20%), RMSE increases to more than 6.15 (+1.65%) for *WS* and 6.56 (+8.43%) for $diff_T$. For LHF, a similar sensitivity pattern is observed, with *WS* and $diff_Q$ showing the largest impacts on model performance, whereas perturbations in other inputs lead to RMSE changes typically within 1%. For β, perturbations in most inputs result in negligible changes (RMSE ranges from 0.22 to 0.23). However, under perturbations of $diff_Q$, the RMSE of β exhibits larger variability due to extreme β estimates, reflecting the ratio-based nature of β and further highlighting the importance of imposing a reasonable β constraint.

We have added this sensitivity analysis into the second paragraph of Section 3.5.3 in the revised manuscript and included the detailed results in Table S10.

"Beyond the sensitivity to the imposed β range, we further examined the robustness of the BrTHF model to uncertainties in its environmental inputs, using one-at-a-time perturbation experiments (Table S10). The results indicate that the model is generally insensitive to perturbations (from ±5 to ±20%) in most auxiliary inputs, with changes in SHF and LHF RMSE typically remaining within 1%. Noticeable sensitivities are mainly associated with *WS*, $diff_Q$, and $diff_T$, which are physically expected given their direct roles in controlling air–sea turbulent heat exchanges. Even under extreme perturbations, the resulting variations in SHF and LHF remain within physically reasonable ranges, suggesting that the BrTHF model does not rely excessively on precise tuning of individual inputs. For β, most perturbations lead to minor changes; however, perturbations of $diff_Q$ can induce large variability due to the ratio-based nature of β, further highlighting the necessity of the β constraint.”

Table S10. Sensitivity of the BrTHF model to different environmental inputs. RMSEs of SHF, LHE, and β under different control scenarios are evaluated against daily observations from 197 buoys, with relative changes compared to the baseline scenario (SHF = 6.05 W/m^2, LHE = 23.67 W/m^2, β = 0.22) shown in parentheses.

Input		**SHF (W/m^2)**	**LHF (W/m^2)**	**β**	**Input**		**SHF (W/m^2)**	**LHF (W/m^2)**	**β**
SLP	**+5%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)	***WS***	**+5%**	6.1 (+0.83%)	23.77 (+0.42%)	0.22 (+0.0%)
	-5%	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)		**-5%**	6.04 (-0.17%)	23.7 (+0.13%)	0.22 (+0.0%)
	+10%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**+10%**	6.17 (+1.98%)	24.01 (+1.44%)	0.22 (+0.0%)
	-10%	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)		**-10%**	6.05 (+0.0%)	23.87 (+0.84%)	0.22 (+0.0%)
	+20%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**+20%**	6.4 (+5.79%)	24.88 (+5.11%)	0.23 (+4.55%)
	-20%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**-20%**	6.15 (+1.65%)	24.61 (+3.97%)	0.22 (+0.0%)
SW	**+5%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)	***SWH***	**+5%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)
	-5%	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)		**-5%**	6.06 (+0.17%)	23.66 (-0.04%)	0.22 (+0.0%)
	+10%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**+10%**	6.06 (+0.17%)	23.68 (+0.04%)	0.22 (+0.0%)
	-10%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**-10%**	6.06 (+0.17%)	23.66 (-0.04%)	0.22 (+0.0%)
	+20%	6.08 (+0.5%)	23.68 (+0.04%)	0.22 (+0.0%)		**+20%**	6.06 (+0.17%)	23.71 (+0.17%)	0.22 (+0.0%)
	-20%	6.07 (+0.33%)	23.68 (+0.04%)	0.22 (+0.0%)		**-20%**	6.07 (+0.33%)	23.68 (+0.04%)	0.22 (+0.0%)
LW	**+5%**	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)	***T_p***	**+5%**	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)

	-5%	6.06 (+0.17%)	23.66 (-0.04%)	0.22 (+0.0%)		**-5%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)
	+10%	6.05 (+0.0%)	23.69 (+0.08%)	0.22 (+0.0%)		**+10%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)
	-10%	6.07 (+0.33%)	23.67 (+0.0%)	0.22 (+0.0%)		**-10%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)
	+20%	6.06 (+0.17%)	23.72 (+0.21%)	0.22 (+0.0%)		**+20%**	6.06 (+0.17%)	23.68 (+0.04%)	0.23 (+4.55%)
	-20%	6.09 (+0.66%)	23.69 (+0.08%)	0.23 (+4.55%)		**-20%**	6.06 (+0.17%)	23.68 (+0.04%)	0.22 (+0.0%)
SSS	**+5%**	6.05 (+0.0%)	23.67 (+0.0%)	0.22 (+0.0%)	**$diff_T$**	**+5%**	6.15 (+1.65%)	23.68 (+0.04%)	0.23 (+4.55%)
	-5%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**-5%**	6.05 (+0.0%)	23.66 (-0.04%)	0.22 (+0.0%)
	+10%	6.05 (+0.0%)	23.68 (+0.04%)	0.22 (+0.0%)		**+10%**	6.33 (+4.63%)	23.7 (+0.13%)	0.23 (+4.55%)
	-10%	6.06 (+0.17%)	23.68 (+0.04%)	0.22 (+0.0%)		**-10%**	6.14 (+1.49%)	23.65 (-0.08%)	0.22 (+0.0%)
	+20%	6.06 (+0.17%)	23.72 (+0.21%)	0.22 (+0.0%)		**+20%**	6.94 (+14.71%)	23.75 (+0.34%)	0.23 (+4.55%)
	-20%	6.08 (+0.5%)	23.71 (+0.17%)	0.22 (+0.0%)		**-20%**	6.56 (+8.43%)	23.65 (-0.08%)	0.22 (+0.0%)
ADT	**+5%**	6.06 (+0.17%)	23.68 (+0.04%)	0.22 (+0.0%)	**$diff_Q$**	**+5%**	6.07 (+0.33%)	23.75 (+0.34%)	1.84 (+736.36%)
	-5%	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)		**-5%**	6.04 (-0.17%)	23.77 (+0.42%)	0.23 (+4.55%)
	+10%	6.06 (+0.17%)	23.7 (+0.13%)	0.23 (+4.55%)		**+10%**	6.1 (+0.83%)	24 (+1.39%)	0.28 (+27.27%)
	-10%	6.07 (+0.33%)	23.7 (+0.13%)	0.22 (+0.0%)		**-10%**	6.04 (-0.17%)	24.08 (+1.73%)	0.23 (+4.55%)
	+20%	6.09 (+0.66%)	23.8 (+0.55%)	0.23 (+4.55%)		**+20%**	6.17 (+1.98%)	24.93 (+5.32%)	2.09 (+850.0%)
	-20%	6.11 (+0.99%)	23.78 (+0.46%)	0.22 (+0.0%)		**-20%**	6.05 (+0.0%)	25.31 (+6.93%)	0.23 (+4.55%)
CS	**+5%**	6.06 (+0.17%)	23.67 (+0.0%)	0.22 (+0.0%)					
	-5%	6.05 (+0.0%)	23.66 (-0.04%)	0.22 (+0.0%)					
	+10%	6.06 (+0.17%)	23.68 (+0.04%)	0.22 (+0.0%)					
	-10%	6.05 (+0.0%)	23.66 (-0.04%)	0.22 (+0.0%)					
	+20%	6.06 (+0.17%)	23.69 (+0.08%)	0.22 (+0.0%)					

-20%	6.05 (+0.0%)	23.66 (-0.04%)	0.22 (+0.0%)

I received the following comment:

"The buoy fluxes are calculated from the meteorological variables aggregated to daily averages which will bias the output compared to native resolution due to the non-linearity of the bulk formulae. There is no information as to whether the cool-skin and warm-layer adjustments are switched on or off in COARE3.5.

Re: Thank you for your comments. We acknowledge that calculating air-sea turbulent heat fluxes from daily-averaged meteorological variables may introduce biases due to the nonlinearity of bulk aerodynamic formulations. In this study, daily buoy-derived air-sea turbulent heat fluxes were computed using daily-averaged meteorological inputs and the COARE3.5 model, with the cool-skin and warm-layer parameterizations switched off. This configuration follows the practice adopted by Pacific Marine Environmental Laboratory for producing daily air-sea turbulent heat flux products at TAO/TRITON, PIRATA, and RAMA buoy sites (https://www.pmel.noaa.gov/tao/drupal/flux/documentation-lw.html), which have been widely used for flux product evaluation (Bourras, 2006).

To assess the potential impact of this temporal aggregation, we conducted an additional analysis by recalculating air-sea turbulent heat fluxes using the original high-frequency buoy observations. For this analysis, the cool-skin and warm-layer parameterizations were enabled only when the required forcings—including a first estimate of the full net surface heat flux and radiative components—were available, consistent with the recommendations of Cronin et al. (2006). These high-frequency flux estimates were then aggregated to daily means (hourly or sub-hourly observations exceeded 80% on a given day) and used to evaluate the performance of the flux products and the retrained BrTHF model.

The comparison indicates that while absolute accuracy metrics differ between the two strategies, the relative performance among different products and models remains unchanged (BrTHF outperformed the other seven products, with RMSE of 6.07 W/m^2, 23.61 W/m^2 and 0.21 for SHF, LHF and β, respectively). Therefore, the main conclusions regarding the comparative performance of BrTHF are not materially affected by the choice of daily-averaged versus high-frequency flux calculations. For consistency with commonly used daily buoy flux datasets, we retain the results based on daily-averaged inputs, while providing the evaluation results based on high-frequency flux calculation in Table S11. Corresponding clarifications have been added to the fouth paragraph of Section 2.1 and fourth paragraph of Section 3.5.3 of the revised manuscript.

"After the above mentioned data preprocessing, the daily buoy-derived air-sea turbulent heat fluxes (SHF and LHF) were then calculated using the daily oceanic and atmospheric measurements combined with the version 3.5 of Coupled Ocean-Atmosphere Response Experiment (COARE3.5) model (Edson, 2013) (available at https://github.com/NOAA-PSL/COARE-algorithm) with the cool-skin and warm-layer calculation switched off. The configuration follows the practice adopted by Pacific Marine Environmental Laboratory for producing daily air-sea turbulent heat flux products (https://www.pmel.noaa.gov/tao/drupal/flux/documentation-lw.html) at Global Tropical Moored Buoy Array (TAO/TRITON, PIRATA, and RAMA)."

"Finally, we examined the potential impact of temporal aggregation, as the buoy-derived daily fluxes used for model training and evaluation were calculated from daily-averaged meteorological variables via COARE3.5, which may introduce biases due to the nonlinearity of bulk flux formulations. By recalculating fluxes from high-frequency buoy observations and aggregating to daily means, we found that although absolute errors differ to some extent, the relative performance among different flux products and models remains unchanged, with the BrTHF model still achieving the best overall accuracy (Table S11). This confirms that our conclusions are robust to the choice of daily flux calculation strategy."

Table S11. Statistical metrics of the spatial ten-fold cross-validation of estimated daily SHF, LHF and β from the BrTHF model and seven widely used products, with and without removing estimated β ($\beta < -5$ or $\beta > 5$ in the seven widely used products) that deviates from the range of β observations collected from the 197 buoys. The observations of SHF, LHF and β

were calculated from the high-frequency meteorological variables at sites.

		All Samples (462082)			Samples (458233) excluding $\beta < -5$ or $\beta > 5$		
		SHF	**LHF**	β	**SHF**	**LHF**	β
JOFURO3	**BIAS (W/m^2)**	2.06	-0.61	0.04	2.01	-0.59	0.03
	RMSE (W/m^2)	9.02	30.32	9.97	8.94	30.36	0.23
	r	0.87	0.86	0.0	0.86	0.85	0.28
IFREMER	**BIAS (W/m^2)**	-4.49	4.74	-0.04	-4.5	4.79	-0.04
	RMSE (W/m^2)	9.8	28.92	5.36	9.76	28.91	0.18
	r	0.91	0.86	0.03	0.91	0.86	0.45
SeaFlux	**BIAS (W/m^2)**	0.48	-2.88	0.01	0.48	-2.87	0.01
	RMSE (W/m^2)	6.92	27.56	0.87	6.88	27.62	0.18
	r	0.9	0.88	0.08	0.9	0.87	0.34
ERA5	**BIAS (W/m^2)**	-2.62	-10.72	-0.01	-2.69	-10.78	-0.01
	RMSE (W/m^2)	7.15	28.86	2.29	7.11	28.95	0.17
	r	0.91	0.89	0.03	0.91	0.89	0.4
MERRA2	**BIAS (W/m^2)**	0.01	-15.5	-0.00	-0.0	-15.6	0.01
	RMSE (W/m^2)	6.53	35.9	8.02	6.5	36.02	0.17
	r	0.92	0.84	0.01	0.91	0.84	0.4
OAFlux	**BIAS (W/m^2)**	-0.02	3.72	0.01	-0.06	3.79	0.01
	RMSE (W/m^2)	7.86	32.09	1.78	7.81	32.16	0.19
	r	0.9	0.83	0.02	0.9	0.83	0.35
OHF	**BIAS (W/m^2)**	3.02	7.74	0.01	2.95	7.81	0.03
	RMSE (W/m^2)	12.11	34.81	23.11	11.74	34.81	0.26
	r	0.69	0.81	0.0	0.71	0.8	0.13
BrTHF	**BIAS (W/m^2)**	-0.15	-0.75	0.01	-0.13	-0.65	0.0
	RMSE (W/m^2)	6.07	23.61	0.21	6.02	23.62	0.15
	r	0.93	0.91	0.26	0.93	0.91	0.42

The differences between the buoy estimates and the flux products will include a contribution from the difference in the parameterisations used in the products and COARE3.5 (see e.g. https://doi.org/10.3389/fmars.2022.1049168 and https://doi.org/10.1175/JPO-D-16-0169.1) so is not really a fair comparison. It would also have been interesting to see a comparison with a traditional implementation of the bulk formulae with the input variables use to force the neural network." > this comment should be addressed.

Re: Thank you for your comments. We agree that differences between the buoy-derived fluxes and existing flux products may partly arise from discrepancies in the underlying parameterisations, which indeed makes such comparisons not strictly method-consistent. To better address this concern, we conducted an additional baseline experiment using the COARE3.5 model forced by a subset of the core daily meteorological variables used in the BrTHF model (e.g., air temperature, humidity, wind speed, and sea surface temperature) with the cool-skin and warm-layer parameterizations switched off.

The RMSEs reach approximately 10.3 W/m^2 for SHF and 34.4 W/m^2 for LHF, accompanied by systematic underestimations of about 6.5 W/m^2 and 21.3 W/m^2, respectively. In addition, physically unrealistic values of β emerge in the COARE3.5 estimates, leading to a degradation in β accuracy (RMSE = 6.35). These results suggest that, when driven by the same set of daily-mean meteorological inputs, the BrTHF model provides more robust estimates of SHF, LHF, and β. In particular, the BrTHF model maintains relatively high estimation accuracy even under conditions where the input meteorological variables contain larger uncertainties.

The corresponding analysis has been added to the third paragraph of Section 3.5.3, with detailed results presented in Table S12.

"To further address potential methodological inconsistencies between buoy-derived fluxes and flux products, we conducted an additional baseline experiment. Specifically, the COARE3.5 model was forced with the same subset of daily meteorological variables used to drive the BrTHF model, thereby providing a method-consistent reference under identical forcing conditions. The results (Table S12) show that the COARE3.5-driven estimates exhibit substantially larger errors for SHF and LHF, with RMSEs of approximately 10.3 W/m^2 and 34.4 W/m^2, respectively, accompanied by systematic underestimations relative to buoy observations. More importantly, physically unrealistic values of the β emerge in the COARE3.5 estimates, leading to a degradation in β accuracy (RMSE = 6.35). These findings suggest that, even when driven by the same meteorological inputs, traditional bulk formulations remain highly sensitive to forcing uncertainties. In contrast, the BrTHF model demonstrates improved robustness by explicitly incorporating physical constraints within the machine-learning framework."

Table S12. Statistical metrics of the spatial ten-fold cross-validation of estimated daily SHF,

LHF and β from the BrTHF model and COARE3.5 model, with and without removing estimated β (β < -5 or β > 5 in the COARE3.5 model) that deviates from the range of β observations collected from the 197 buoys.

		All Samples (463585)			Samples (463127) excluding $\beta < -5$ or $\beta > 5$		
		SHF	LHF	β	SHF	LHF	β
BrTHF	**BIAS (W/m^2)**	0.09	0.14	-0.01	0.13	0.08	0.00
	RMSE (W/m^2)	6.05	23.67	0.22	6.02	23.7	0.21
	r	0.93	0.91	0.25	0.93	0.91	0.31
COARE3.5	**BIAS (W/m^2)**	-6.48	-21.32	-0.03	-6.49	-21.34	-0.03
	RMSE (W/m^2)	10.3	34.44	6.35	10.3	34.45	0.23
	r	0.9	0.9	0.01	0.9	0.9	0.28

Reference:

Bourras, D.: Comparison of five satellite-derived latent heat flux products to moored buoy data, Journal of Climate, 19, 6291-6313, 2006.

Cronin, M. F., Fairall, C. W., and McPhaden, M. J.: An assessment of buoy-derived and numerical weather prediction surface heat fluxes in the tropical Pacific, Journal of Geophysical Research: Oceans, 111, 10.1029/2005jc003324, 2006.

Edson, J. B. a. J., Venkata and Weller, Robert A and Bigorre, Sebastien P and Plueddemann, Albert J and Fairall, Christopher W and Miller, Scott D and Mahrt, Larry and Vickers, Dean and Hersbach, Hans: On the Exchange of Momentum over the Open Ocean, Journal of Physical Oceanography, 43, 1589-1610, 10.1175/jpo-d-12-0173.1, 2013.

Jo, Y.-H.: Calculation of the Bowen ratio in the tropical Pacific using sea surface temperature data, Journal of Geophysical Research, 107, 10.1029/2001jc001150, 2002.

Wang, W., Chakraborty, T. C., Xiao, W., and Lee, X.: Ocean surface energy balance allows a constraint on the sensitivity of precipitation to global warming, Nat Commun, 12, 2115, 10.1038/s41467-021-22406-7, 2021.

Yang, Y., Sun, H., Wang, J., Zhang, W., Zhao, G., Wang, W., Cheng, L., Chen, L., Qin, H., and Cai, Z.: Global ocean surface heat fluxes derived from the maximum entropy production framework accounting for ocean heat storage and Bowen ratio adjustments, Earth System Science Data, 17, 1191-1216, 10.5194/essd-17-1191-2025, 2025.