Review 1

This is a novel approach to calculate turbulent heat fluxes over the ocean. Aside from some usual manuscript omissions and corrections (noted below), my main concern is on the description and complete understanding by the authors of the data sets used, in particular, the SSM/I products that they presumably obtained from RSS. In my opinion, they lack full understanding of this data set, what the limitations may be, etc. but use them on face value. Here are just some of my concerns:

Response:

We thank the reviewer for the constructive comments. We fully agree that our original manuscript did not sufficiently document the exact SSM/I–SSMIS dataset, versioning, intercalibration, constellation sampling characteristics, and limitations. In the revision, we substantially expand the SSM/I dataset description, add formal dataset citations, and explicitly discuss orbit-time drift, multi-satellite overlaps, and implications for long-term analyses (Section 2.1.1). The revised text is supported by RSS technical documentation and calibration reports.The main text has been amended as follows:

"2.1.1 SSM/I data

The Special Sensor Microwave/Imager (SSM/I), flown on the Defense Meteorological Satellite Program (DMSP) series, is a conically scanning passive microwave radiometer designed to measure naturally emitted microwave radiation from Earth's surface and atmosphere. Since its initial deployment in 1987, SSM/I has provided synoptic, near-all-weather observations widely used for both operational weather applications and climate studies (Hollinger et al., 1990). The instrument carries seven frequency channels (19.35–85.5 GHz) that enable the retrieval of a variety of geophysical parameters. DMSP platforms follow near-polar orbits and typically provide two passes per day (ascending and descending), but the local equator-crossing time differs among satellites and drifts over mission life because the orbits are not maintained in strict local-time control (Wentz, 2013). In addition, multiple DMSP satellites often operate concurrently, increasing sampling but also introducing local-time heterogeneity in the long-term record (Fennig et al., 2020). Beginning in 2003, the Special Sensor Microwave Imager/Sounder (SSMIS) replaced SSM/I, adding sounding channels and extending high-frequency capabilities, thereby enhancing precipitation and cloud microphysical retrievals (Bommarito, 1993). Because of orbital geometry and conical scanning, swath gaps remain between adjacent passes, particularly in the tropics.

In this study, we use four key ocean atmosphere variables retrieved from the SSM/I to SSMIS series SSW, CLW, total column WV, and RR—together with SST from NOAA OISST v2.1 to retrieve global near-surface T_a *and* Q_a*. Specifically, we use the Remote*

Sensing Systems (RSS) SSM/I to SSMIS Ocean Products (daily gridded fields; Version 7) for DMSP satellites F10-F17 over January 1992 December 2020 (RSS, 2025b; Wentz, 2013). RSS applies a unified physically based retrieval algorithm and Version-7 calibration/intercalibration to promote consistency across satellites and the SSM/I to SSMIS transition (Wentz, 2013; RSS, 2025b). We process ascending and descending passes separately and provide both pass-time-resolved daily fields in the final product.

Despite these processing efforts, certain data limitations remain. Swath gaps and rain contamination lead to missing or degraded retrievals in some regions and time periods, and local sampling times vary and drift across the DMSP constellation (Wentz, 2013; RSS, 2025a). To minimize local-time aliasing, we (I) treat ascending and descending branches separately throughout the gap-filling and retrieval procedures (Section 2.2) and (II) recommend using the mean of the two branches for long-term trend analyses (Section 5), while retaining each branch for studies focused on pass-time (diurnal sampling) characteristics."

I dont even see a citation to the SSM/I data set used, aside from the very end.

Response:

Agreed. We now formally cite the RSS SSM/I–SSMIS Ocean Products (Version 7) in the data section (not only in acknowledgements), and add the RSS mission documentation and Version-7 calibration report as references.The main text has been amended as follows:

"In this study, we use four key ocean atmosphere variables retrieved from the SSM/I to SSMIS series SSW, CLW, total column WV, and RR—together with SST from NOAA OISST v2.1 to retrieve global near-surface T_a *and* Q_a*. Specifically, we use the Remote Sensing Systems (RSS) SSM/I to SSMIS Ocean Products (daily gridded fields; Version 7) for DMSP satellites F10-F17 over January 1992 December 2020 (RSS, 2025b; Wentz, 2013). RSS applies a unified physically based retrieval algorithm and Version-7 calibration/intercalibration to promote consistency across satellites and the SSM/I to SSMIS transition (Wentz, 2013; RSS, 2025b). We process ascending and descending passes separately and provide both pass-time-resolved daily fields in the final product.*

Despite these processing efforts, certain data limitations remain. Swath gaps and rain contamination lead to missing or degraded retrievals in some regions and time periods, and local sampling times vary and drift across the DMSP constellation (Wentz, 2013; RSS, 2025a). To minimize local-time aliasing, we (I) treat ascending and descending branches separately throughout the gap-filling and retrieval procedures (Section 2.2) and (II) recommend using the mean of the two branches for long-term trend analyses (Section 5), while retaining each branch for studies focused on pass-time (diurnal sampling) characteristics."

There are numerous SSM/I data sets available, why was this one selected?

Response:

We selected RSS V7 because it provides a long, internally consistent, multi-satellite SSM/I–SSMIS record produced with a unified physically based retrieval algorithm and documented cross-sensor calibration/intercalibration.We have explained the reasons in the main text：

"The Special Sensor Microwave/Imager (SSM/I), flown on the Defense Meteorological Satellite Program (DMSP) series, is a conically scanning passive microwave radiometer designed to measure naturally emitted microwave radiation from Earth's surface and atmosphere. Since its initial deployment in 1987, SSM/I has provided synoptic, near-all-weather observations widely used for both operational weather applications and climate studies (Hollinger et al., 1990). The instrument carries seven frequency channels (19.35–85.5 GHz) that enable the retrieval of a variety of geophysical parameters. DMSP platforms follow near-polar orbits and typically provide two passes per day (ascending and descending), but the local equator-crossing time differs among satellites and drifts over mission life because the orbits are not maintained in strict local-time control (Wentz, 2013). In addition, multiple DMSP satellites often operate concurrently, increasing sampling but also introducing local-time heterogeneity in the long-term record (Fennig et al., 2020). Beginning in 2003, the Special Sensor Microwave Imager/Sounder (SSMIS) replaced SSM/I, adding sounding channels and extending high-frequency capabilities, thereby enhancing precipitation and cloud microphysical retrievals (Bommarito, 1993). Because of orbital geometry and conical scanning, swath gaps remain between adjacent passes, particularly in the tropics.

In this study, we use four key ocean atmosphere variables retrieved from the SSM/I to SSMIS series SSW, CLW, total column WV, and RR—together with SST from NOAA OISST v2.1 to retrieve global near-surface T_a *and* Q_a*. Specifically, we use the Remote Sensing Systems (RSS) SSM/I to SSMIS Ocean Products (daily gridded fields; Version 7) for DMSP satellites F10-F17 over January 1992 December 2020 (RSS, 2025b; Wentz, 2013). RSS applies a unified physically based retrieval algorithm and Version-7 calibration/intercalibration to promote consistency across satellites and the SSM/I to SSMIS transition (Wentz, 2013; RSS, 2025b). We process ascending and descending passes separately and provide both pass-time-resolved daily fields in the final product."*

What intercalibration did RSS employ to allow for a seemless transition between satellites and SSM/I to SSMIS?

Response:

We now summarize RSS Version-7 calibration and intercalibration. RSS V7 aims to apply a calibration methodology consistently across satellite microwave imagers and ties the record to a common reference, enabling continuity between SSM/I and its follow-on SSMIS.

To ensure a seamless transition and establish a consistent Fundamental Climate Data Record (FCDR) between the DMSP SSM/I and SSMIS sensors, Remote Sensing Systems (RSS) employed a rigorous intercalibration methodology centered on a physical Radiative Transfer Model (RTM) combined with a double-difference technique.

First, RSS applied necessary sensor-level corrections to address SSMIS-specific hardware characteristics, such as the emissive antenna effect and solar or lunar intrusions into the calibration targets. Following this, an RTM was utilized to simulate theoretical Top-Of-Atmosphere brightness temperatures over highly stable Earth reference targets.

By comparing the observed brightness temperatures of both SSM/I and SSMIS against these RTM-simulated baselines, RSS calculated the individual sensor residuals. A double-difference approach was then applied across these shared reference targets to determine the relative systematic biases between the respective sensors. These derived offsets were subsequently used to seamlessly align the SSMIS measurements with the SSM/I historical record at the foundational radiance level.

Consequently, the Version-7 data provided by RSS, which we utilized in this study, has already strictly undergone these comprehensive intercalibration procedures. Because our research directly employs this climate-quality, seamlessly transitioned V7 dataset, we did not need to perform additional satellite-to-satellite radiance calibrations in our work. The specific mathematical formulations and empirical coefficients for this intercalibration process are omitted here for brevity, but they are comprehensively detailed in the foundational RSS calibration reports and related literature (Wentz et al., 2013; Meissner & Wentz, 2012).

Starting with F10, DMSP operated two satellites, one around 6 am/6 pm, the other around 10 am/10 pm. You fail to described or recognize this.

Response:

We agree with the reviewer and appreciate the opportunity to clarify this important detail. We are fully aware of the DMSP dual-satellite constellation history starting with F10, and we have accounted for this in our data processing, though we admit we did not explicitly describe the specific orbital times in the original text.

To be specific, we understand that maintaining two operational satellites provided significant temporal overlap. For instance, in 1999, the simultaneous operation of F13 (with approximately 6 am descending and 6 pm ascending overpasses) and F14 (with approximately 10 am descending and 10 pm ascending overpasses) meant that a single grid cell could theoretically be observed up to four times a day (around 6 am, 10 am, 6 pm, and 10 pm).

In our study, integrating data from these concurrently operating satellites is exactly what generates the overlapping records mentioned in our methodology. Because these four potential daily observations capture real sub-daily variability (i.e., local-time heterogeneity), we did not simply average them. Instead, we describe how overlaps are handled in matchup processing (duplicate-record selection). The main body content is as follows:

"Matchup data, which pair satellite retrievals with coincident in situ measurements, are essential for calibrating retrieval algorithms and evaluating data quality. As shown in Table 1 in the second step, Fine-tuning, it is necessary to match SSM/I satellite data with in situ observations to retrieve T_a *and* Q_a*. In this study, we use variables retrieved from the SSM/I satellite, including SSW, CLW, WV, and RR. All satellite data are divided into ascending and descending passes, corresponding to the satellite's northbound and southbound orbits, respectively. We utilize data from DMSP satellites F10-F17, which have overlapping operational periods. Because overlapping DMSP satellites may sample different local times, their concurrent observations can reflect not only sensor differences but also real sub daily variability. This temporal overlap allows simultaneous observations from multiple satellites, as shown in Table 2, thereby increasing data redundancy. For periods with duplicate satellite data, we select the record with the lowest RMSE compared to in situ measurements as the final entry. When in situ data are unavailable for comparison, data from the most recent satellite are retained. This ensures each satellite observation corresponds to one ground-truth measurement. Additionally, 220 days of missing observations are filled using interpolated ERA5 reanalysis data at a 1 ° spatial resolution (Table S1). The final satellite dataset spans 28 years, as detailed in Table 1 of the final step in the Calibration Model. The specific time coverage for each satellite is detailed in Table 3."*

These satellites drift in time, you dont mention this. (you do note in Table 3 the overlap periods, some aspects of the overpass times would be nice to include).

Response:

We thank the reviewer for pointing this out. We acknowledge that the local equator-crossing times of the satellites gradually drift over their mission lifetimes. However, explicitly correcting for this temporal drift is beyond the scope of our current processing framework for two main reasons. First, the foundational RSS Version-7 (V7) dataset used in this study has already undergone rigorous intercalibration to ensure climate-quality consistency across the multi-satellite record. Second, the DeepFlux product is based on daily averages. By aggregating ascending and descending passes and applying an RMSE-based selection strategy during overlapping satellite periods, the potential sampling biases associated with orbital drift are largely smoothed and minimized in the final daily product.

Therefore, these limitations do not affect the main conclusions of this study. The DeepFlux dataset is derived from daily aggregated satellite observations and has been independently validated against buoy measurements.

The main text has been amended as follows:

"Despite these processing efforts, certain data limitations remain. Swath gaps and rain contamination lead to missing or degraded retrievals in some regions and time periods, and local sampling times vary and drift across the DMSP constellation (Wentz, 2013; RSS, 2025a). To minimize local-time aliasing, we (I) treat ascending and descending branches separately throughout the gap-filling and retrieval procedures (Section 2.2) and (II) recommend using the mean of the two branches for long-term trend analyses (Section 5), while retaining each branch for studies focused on pass-time (diurnal sampling) characteristics."

My primary suggestion would be to include more details on this data set and demonstrate your understanding of what you are using. Otherwise, it just seems to be a huge data exercise.

Some general comments (and this is not all of them)

Not all references are in the list and some are out of order. In particular, those from Wang et al.

Response:

We added missing key references for RSS V7 calibration, RSS mission documentation/crossing times, OISST v2.1 improvements, and SST intercomparison/assessment documents, and we will ensure the final reference list orde r and completeness before resubmission.

Figures - I find them extremely crowded and small to see.

Response:

We agree and improve readability by increasing font sizes and splitting the densest multi-panel figures (especially the scatterplot panels) into two figures and/or moving secondary panels to the Supplement.

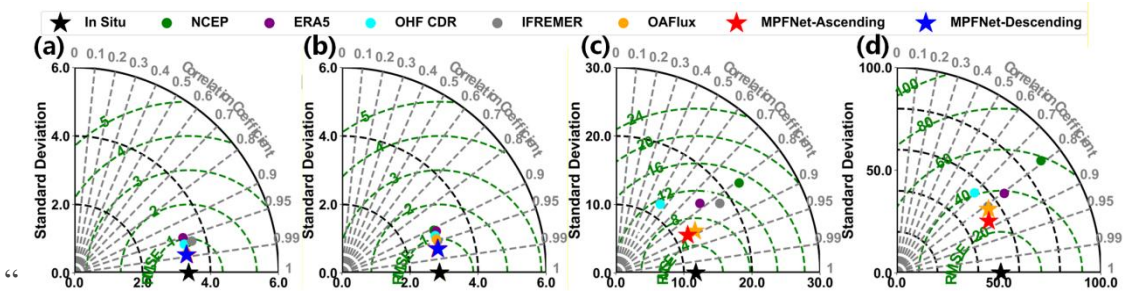


Figure 2. Taylor diagrams comparing NCEP, ERA5, OHF CDR, IFREMER, OAFlux, MPFNet-Ascending, and MPFNet-Descending with in situ observations for (a) T_a, (b) Q_a, (c) SHF, and (d) LHF. The radial distance represents the standard deviation (STD), the azimuthal angle represents the correlation coefficient (CC), and the green dashed contours indicate the centered root-mean-square error (RMSE). In situ observations are denoted by the black star; NCEP, ERA5, OHF CDR, IFREMER, and OAFlux are shown by colored circles; and MPFNet-Ascending and MPFNet-Descending are shown by the red and blue stars, respectively.

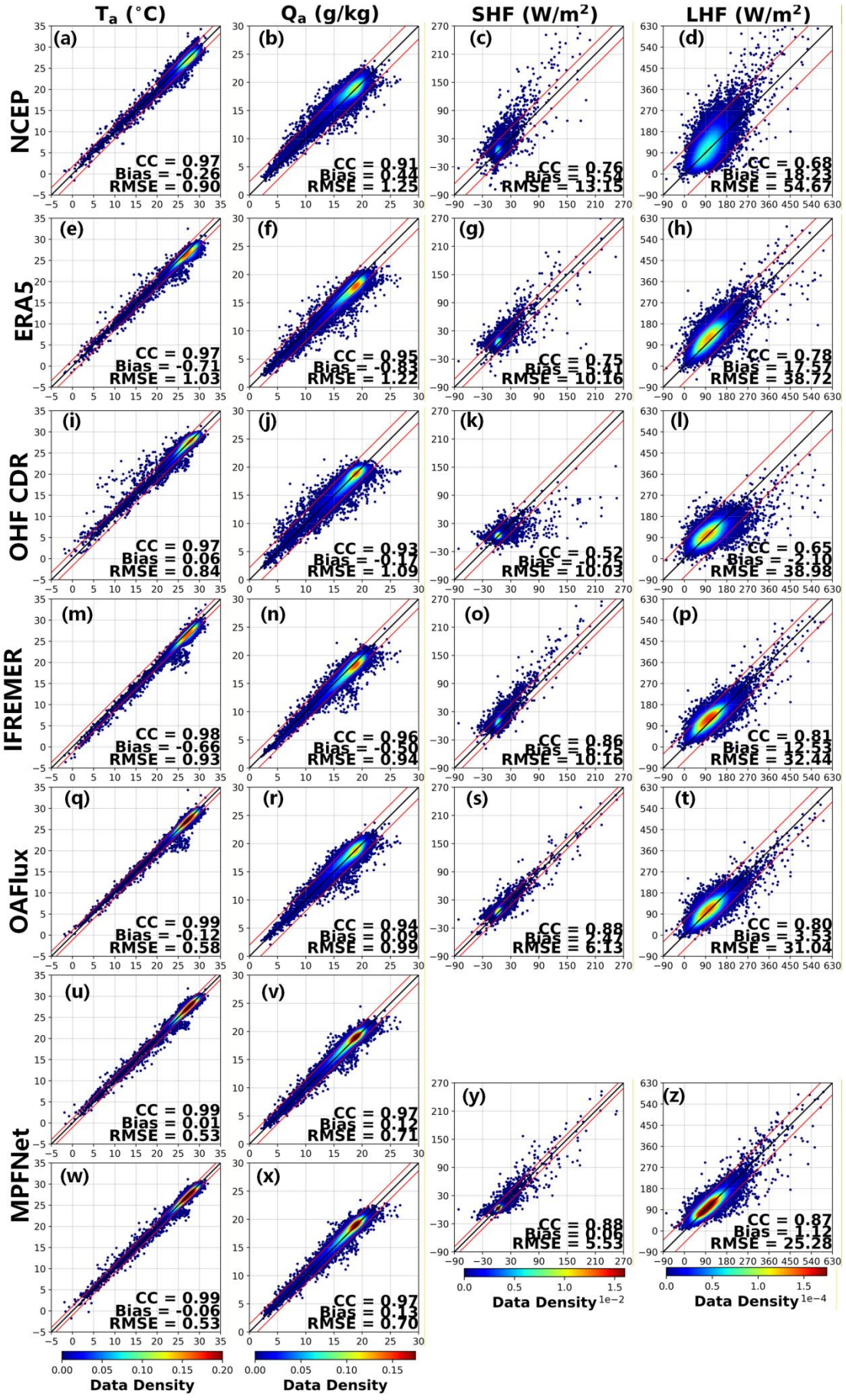

T_a (°C)
Q_a (g/kg)
SHF (W/m^2)
LHF (W/m^2)
NCEP
ERA5
OHF CDR
IFREMER
OAFlux
MPFNet
(a) CC = 0.97 Bias = -0.26 RMSE = 0.90
(b) CC = 0.91 Bias = 0.44 RMSE = 1.25
(c) CC = 0.76 Bias = 5.54 RMSE = 13.15
(d) CC = 0.68 Bias = 18.23 RMSE = 54.67
(e) CC = 0.97 Bias = -0.71 RMSE = 1.03
(f) CC = 0.95 Bias = -0.83 RMSE = 1.22
(g) CC = 0.75 Bias = 5.41 RMSE = 10.16
(h) CC = 0.78 Bias = 17.57 RMSE = 38.72
(i) CC = 0.97 Bias = 0.06 RMSE = 0.84
(j) CC = 0.93 Bias = -0.17 RMSE = 1.09
(k) CC = 0.52 Bias = -0.25 RMSE = 10.03
(l) CC = 0.65 Bias = -2.10 RMSE = 38.98
(m) CC = 0.98 Bias = -0.66 RMSE = 0.93
(n) CC = 0.96 Bias = -0.50 RMSE = 0.94
(o) CC = 0.86 Bias = 6.26 RMSE = 10.16
(p) CC = 0.81 Bias = 12.53 RMSE = 32.44
(q) CC = 0.99 Bias = -0.12 RMSE = 0.58
(r) CC = 0.94 Bias = 0.09 RMSE = 0.99
(s) CC = 0.88 Bias = 2.47 RMSE = 6.13
(t) CC = 0.80 Bias = 3.53 RMSE = 31.04
(u) CC = 0.99 Bias = 0.01 RMSE = 0.53
(v) CC = 0.97 Bias = 0.12 RMSE = 0.71
(w) CC = 0.99 Bias = -0.06 RMSE = 0.53
(x) CC = 0.97 Bias = 0.13 RMSE = 0.70
(y) CC = 0.88 Bias = 0.06 RMSE = 5.53
(z) CC = 0.87 Bias = 1.12 RMSE = 25.28
Data Density
Data Density 1e−2
Data Density 1e−4

Figure 3. Scatterplots of T_a, Q_a, SHF, and LHF retrieved from different products against in situ observations. Panels (a–d), (e–h), (i–l), (m–p), and (q–t) correspond to NCEP, ERA5, OHF CDR, IFREMER, and OAFlux, respectively, for T_a (first column), Q_a (second column), SHF (third column), and LHF (fourth column). Panels (u–v) and (w–x) show the MPFNet results for T_a and Q_a from the two orbital branches (ascending and descending, respectively), while panels (y–z) show the MPFNet estimates of SHF and LHF. The black line denotes the 1:1 reference line, and the two red lines indicate the ±2 standard-deviation envelope of the retrieval–observation differences, encompassing approximately 95% of the samples. Color shading indicates data density, and the CC, Bias, and RMSE values are given in each panel.

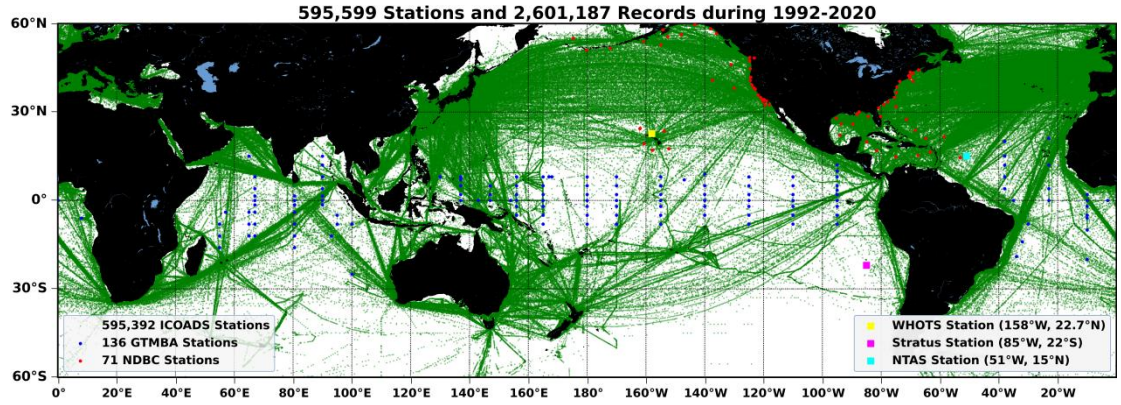


Figure 4. Spatial distribution of the 71 NDBC, 136 GTMBA, and 595,392 ICOADS stations matched with SSM/I during 1992–2020, yielding a total of 2,601,187 satellite–in situ matchup records. The independent validation stations used in this study are highlighted by colored squares: WHOTS (158°W, 22.7°N; yellow), Stratus (85°W, 22°S; magenta), and NTAS (51°W, 15°N; cyan).

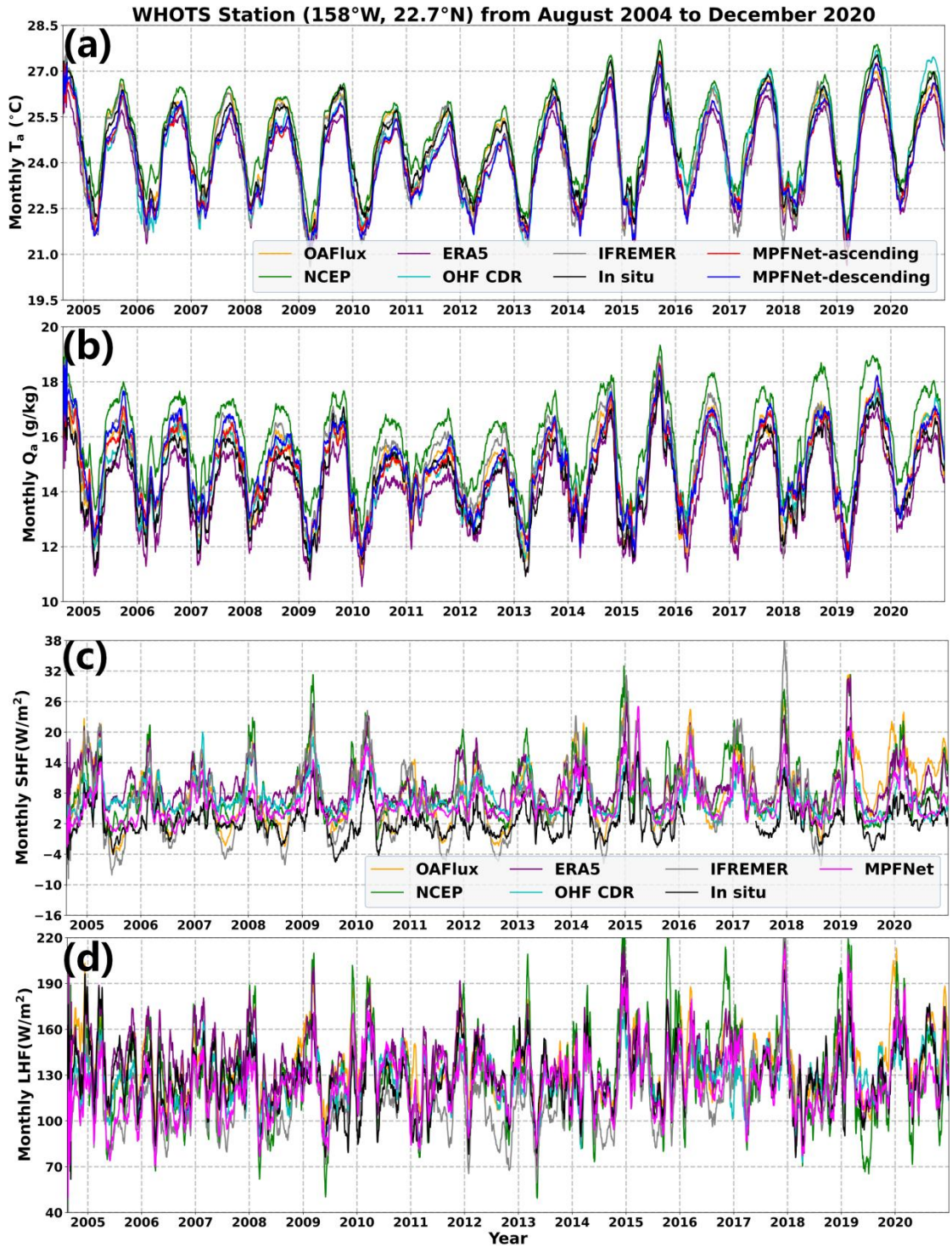


Figure 5. Monthly time series comparison at the WHOTS station (158°W, 22.7°N) from August 2004 to December 2020. Panels (a–d) show T_a, Q_a, SHF, and LHF, respectively. OAFlux, NCEP, ERA5, OHF CDR, IFREMER, and MPFNet retrievals are compared with in situ observations. For T_a and Q_a, the MPFNet ascending and descending retrievals are shown separately, whereas for SHF and LHF the MPFNet estimates are shown as a single series.

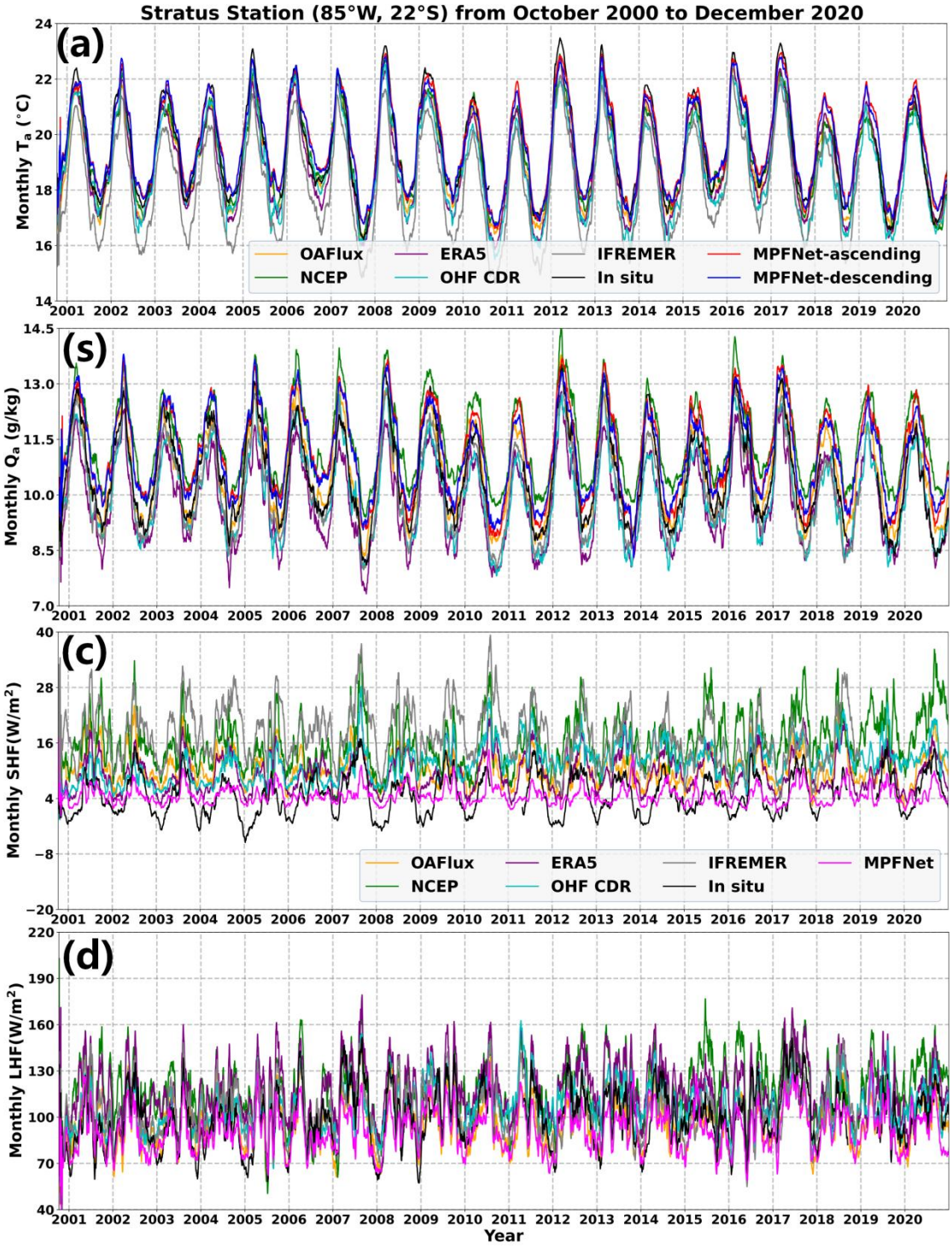


Figure 6. Monthly time series comparison at the Stratus station (85°W, 22°S) from October 2000 to December 2020. Panels (a–d) show T_a, Q_a, SHF, and LHF, respectively. OAFlux, NCEP, ERA5, OHF CDR, IFREMER, and MPFNet retrievals are compared with in situ observations. For T_a and Q_a, the MPFNet ascending and descending retrievals are shown separately, whereas for SHF and LHF the MPFNet estimates are shown as a single series.

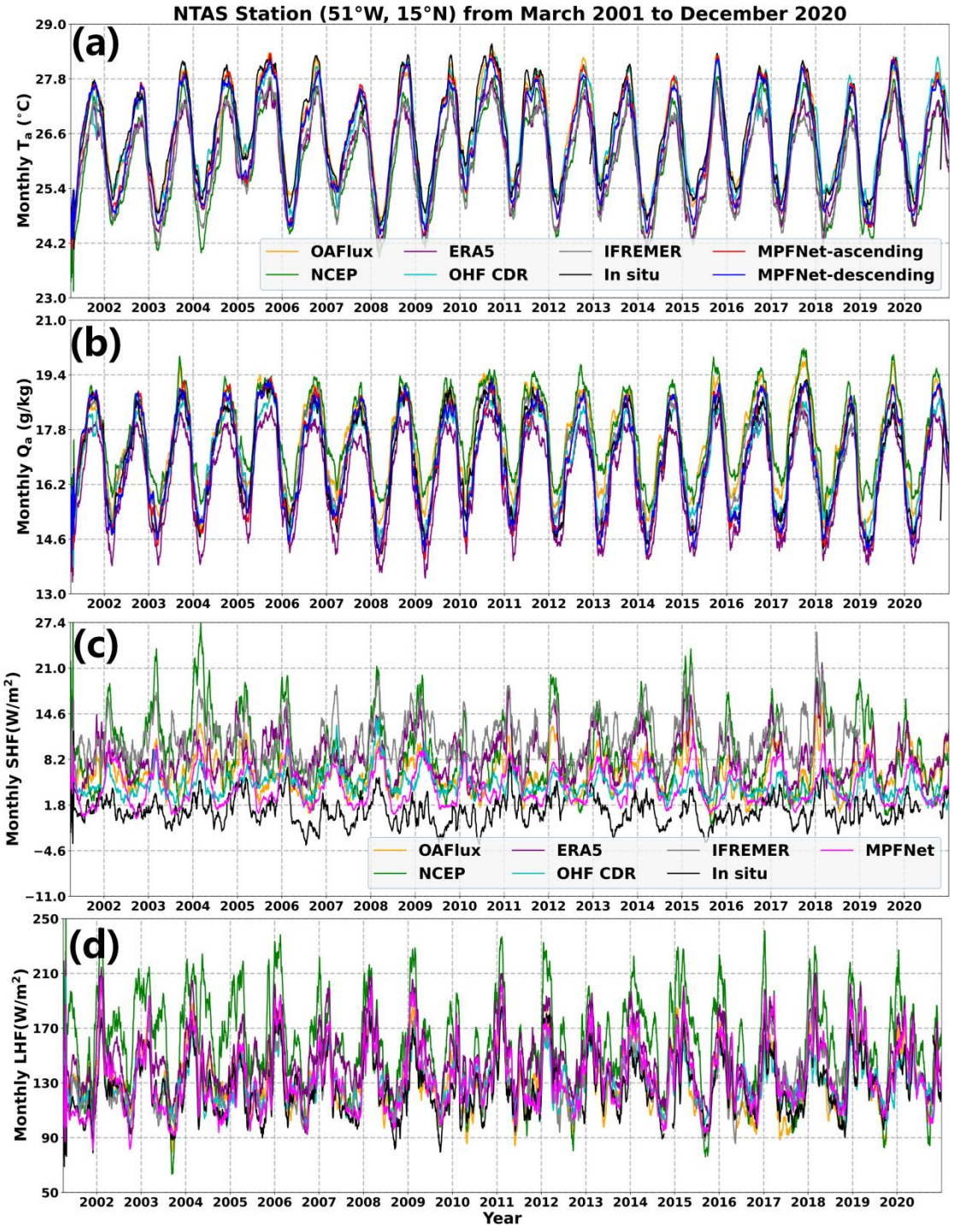


Figure 7. Monthly time series comparison at the NTAS station (51°W, 15°N) from March 2001 to December 2020. Panels (a–d) show T_a, Q_a, SHF, and LHF, respectively. OAFlux, NCEP, ERA5, OHF CDR, IFREMER, and MPFNet retrievals are compared with in situ observations. For T_a and Q_a, the MPFNet ascending and descending retrievals are shown separately, whereas for SHF and LHF the MPFNet estimates are shown as a single series.

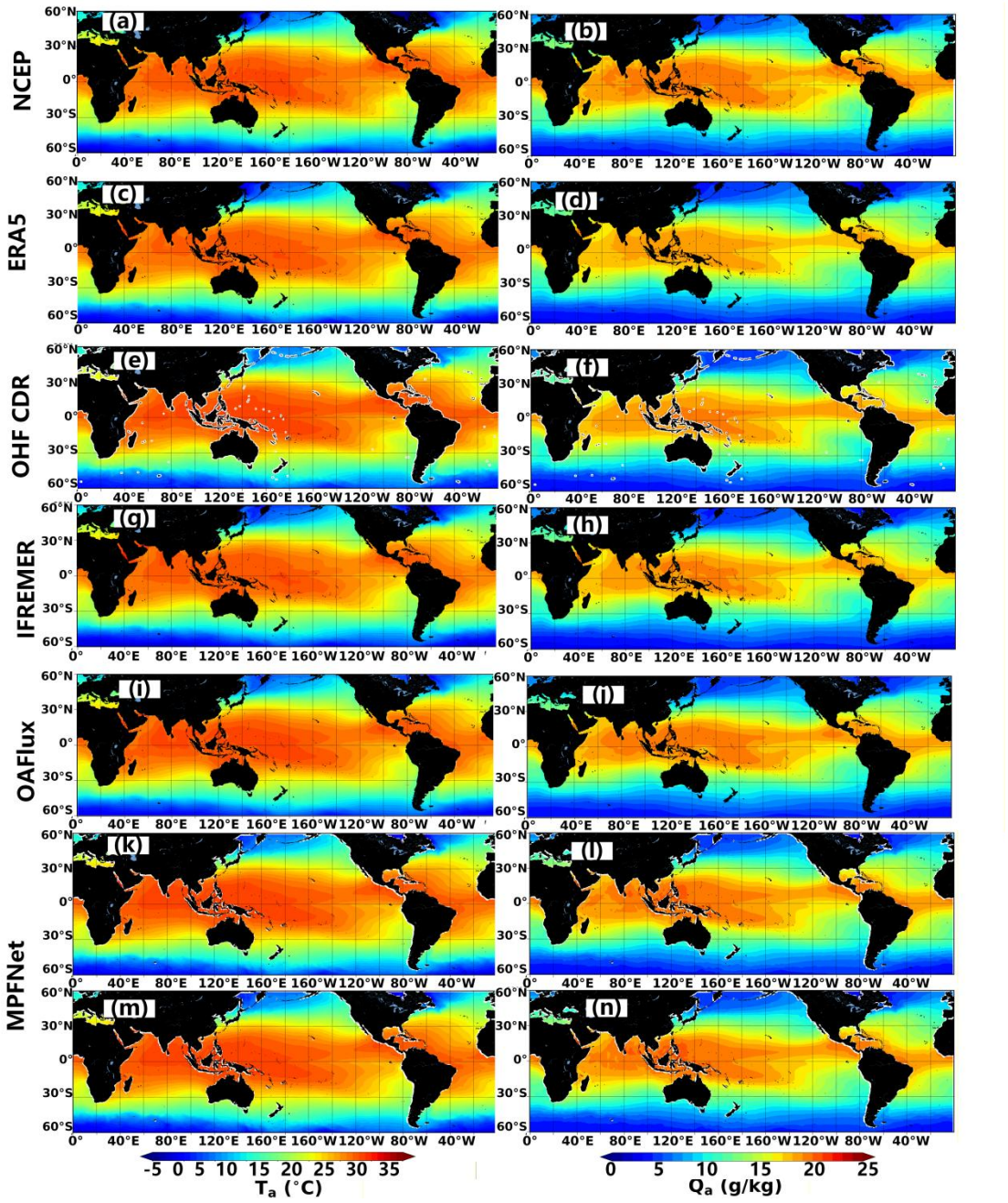


Figure 8. Global spatial distribution of the annual mean T_a and Q_a in 2018 from different products. The left column shows T_a and the right column shows Q_a. Panels (a, b) correspond to NCEP, (c, d) to ERA5, (e, f) to OHF CDR, (g, h) to IFREMER, and (i, j) to OAFlux. Panels (k, l) and (m, n) show the MPFNet retrievals from the ascending and descending orbits, respectively.

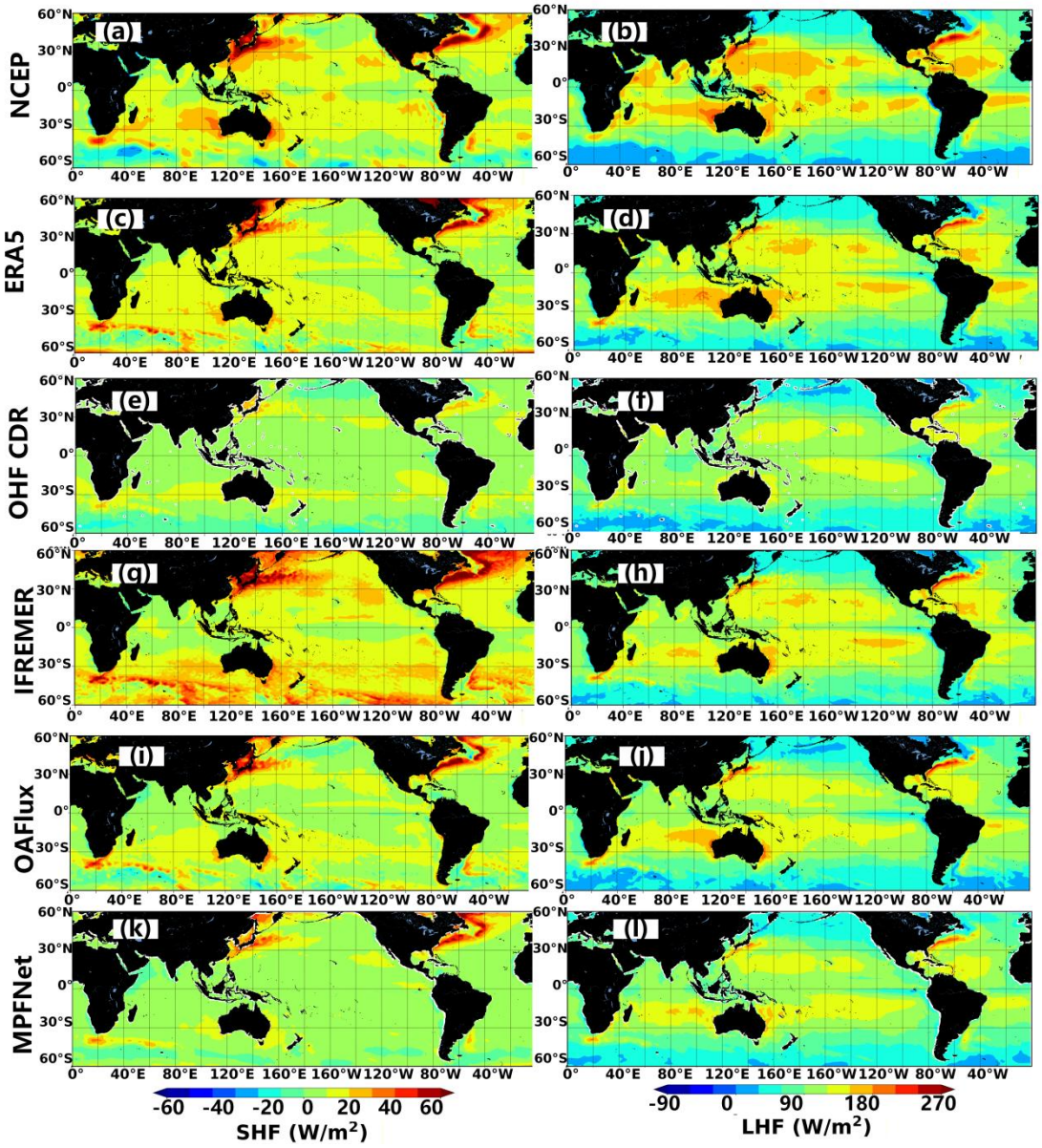


Figure 9. Global spatial distribution of the annual mean SHF and LHF in 2018 from different products. The left column shows SHF and the right column shows LHF. Panels (a, b) correspond to NCEP, (c, d) to ERA5, (e, f) to OHF CDR, (g, h) to IFREMER, (i, j) to OAFlux, and (k, l) to MPFNet.

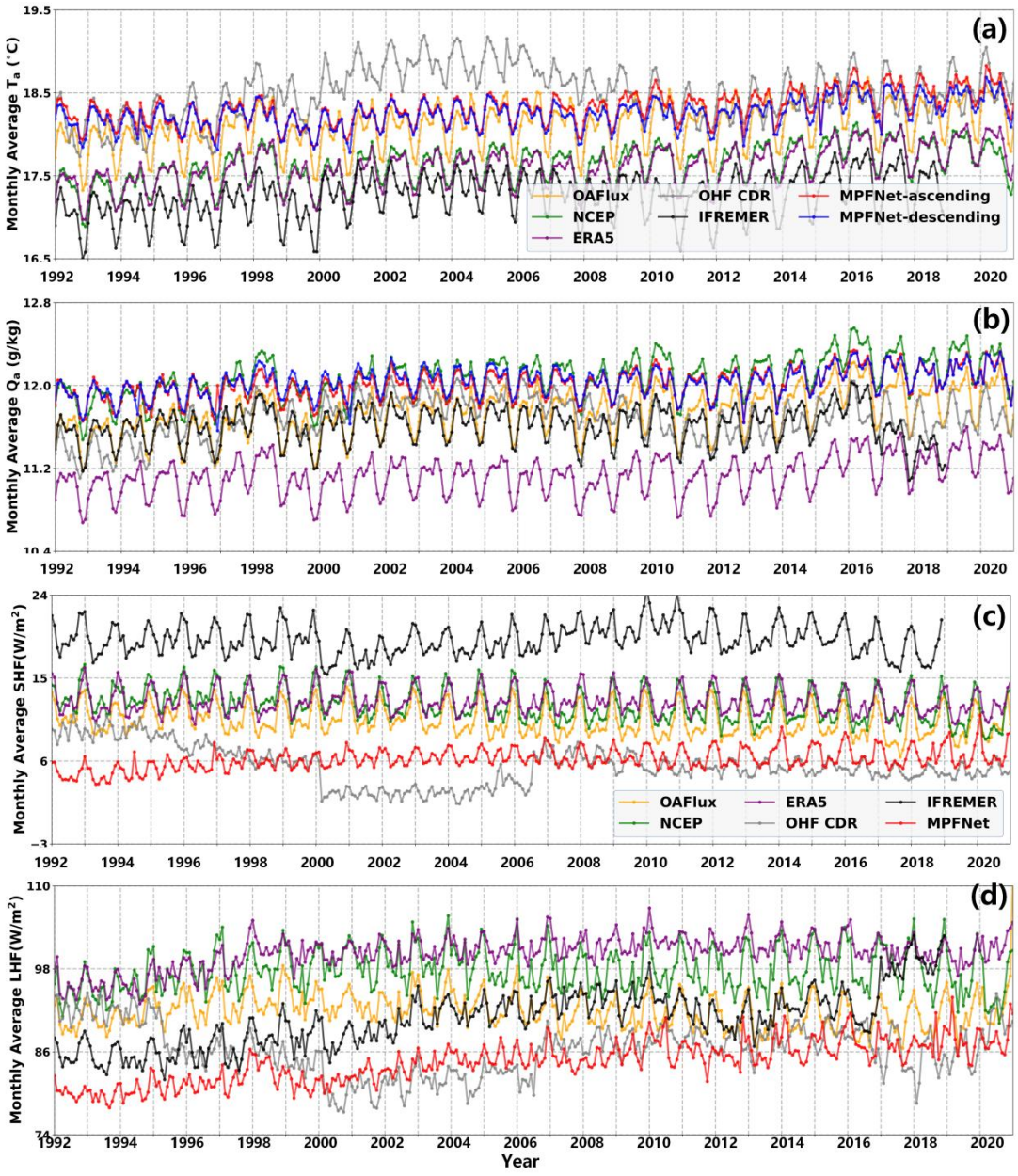


Figure 10. The temporal evolution of the monthly average of global (a) T_a, (b) Q_a, (c) SHF and (d) LHF for different products from 1992 to 2020."

- Line 56 - replace "bottleneck" with "obstacle"

Response:

Corrected.

- Lines 97-100: SSM/II was designed to support US Naval operations on a weather scale. It has become a cornerstone for climate studies but you should be aware of

its weather usage. There are numerous applications - hurricane monitoring, heavy rainfall, etc.

Response:

We revised the SSM/I description to explicitly acknowledge both operational/weather and climate applications, citing an authoritative instrument paper(Wang, 2025a; Wang, 2025b).

Wang, H., Zhou, Y., and Li, X.: GDCM: Generalized data completion model for satellite observations, Remote Sensing of Environment, 324, 114760, 2025.

Wang, M., H. Wang and X. Li, Enhancing Retrievals of Air-Sea Heat Fluxes from AMSR2 Microwave Observations Based on Deep Learning, IEEE Transactions on Geoscience and Remote Sensing, 2025.

- The formatting in Table 1 is poor, its hard to distinguish the various dats sets - could you use a table with gridlines to stratify the information?

Response:

We reformatted Table 1 with clear gridlines and corrected the DeepFlux period to be consistent with the manuscript (1992–2020).

- I would be consistent in defining ascending and descending orbits throughout the paper - using North and South is not commonly used.

Response:

We now consistently use “ascending” and “descending” throughout.

Review 2

This paper presents a method and dataset for global ocean heat flux over an almost 30 year period. The method draws on re-analysis data to machine-learn how to "complete" SSMI-series satellite fields of data. From those completed fields for variables such as surface temperature, humidity and wind speeds, a bulk-formulae based module computes fluxes (one has to read another paper to understand how that step is formulated more fully). In comparison with in situ based measurements of these components and associated fluxes, the authors present results suggesting the new product is scientifically competitive with established products. Some analysis of the drivers of long-term trends in the fluxes of the new product are shown.

It is an interesting contribution to the development of better quantification of air-sea fluxes. My critical comments on the work are as follows.

Response:

We thank the reviewer for their insightful comments on structural uncertainty, independence, and the definition of the measurand. In response, we (i) add sensitivity/structural-uncertainty experiments (Section 4.5), (ii) clarify training/validation independence and explicitly withhold three moored stations for truly independent evaluation, and (iii) clarify that DeepFlux provides pass-time-resolved daily fields (ascending/descending), with recommended usage for trends. We also add an SST trend caveat and a sensitivity test using an alternative SST CDR.

The method of "SSMI completion" is heavily machine-learning led. The approach is presented, inevitably, at a relatively high level. It sounds methodical and reasonable, but nonetheless, with such an approach, many specific design choices are made the affect the result. Other choices could have been made, and this structural uncertainty in the design is not explored. This seems to me to be a general problem with machine-learning approaches, where choices for processing are not really based on physical understanding or hypotheses: the inability to attribute the outcomes to scientific uncertainties, because machine-learning design choices are at least as important.

Response:

We thank the Reviewer for this important and constructive comment. We agree that machine-learning pipelines involve design choices (e.g., input temporal context, training strategy, network components), and these choices can introduce structural uncertainty if not assessed. Our manuscript is primarily a dataset paper, and thus the method description is intentionally concise; however, the key methodological choices and their sensitivity/ablation tests have been comprehensively evaluated in our two companion peer-reviewed methodology papers: (i) the Generalized Data Completion Model (GDCM) for spatio-temporal gap filling, and (ii) the Matrices-Points Fusion Network (MPFNet) for Ta/Qa retrieval and flux estimation.(Wang, 2025a; Wang, 2025b).

Specifically, the GDCM paper reports sensitivity tests on the input temporal length (e.g., 1-day vs 7-day context) and the incremental-learning training strategy designed to gradually introduce realistic missing patterns, showing substantial skill improvements when adopting the selected design.

The MPFNet paper includes explicit ablation studies that quantify the impact of (a) physical predictor selection, (b) matrix size (spatial context), and (c) key model techniques (e.g., FNO, ResNet, and transfer learning with ERA5 pretraining) on retrieval/flux skill, and identifies the final configuration used in this dataset as the optimal choice.

In response to the Reviewer, we revised the manuscript to (1) add a short subsection summarizing these sensitivity/ablation findings and explicitly referencing the two methodology papers, (2) include the key ablation/sensitivity figures (or re-plotted equivalents) in the Supplement, and (3) add a limitations paragraph acknowledging remaining structural uncertainty and outlining future work (e.g., ensemble/alternative architectures to quantify structural uncertainty). These additions clarify the physical/engineering rationale of our design choices and demonstrate that the adopted configurations are supported by quantitative sensitivity evidence.

Wang, H., Zhou, Y., and Li, X.: GDCM: Generalized data completion model for satellite observations, Remote Sensing of Environment, 324, 114760, 2025.

Wang, M., H. Wang and X. Li, Enhancing Retrievals of Air-Sea Heat Fluxes from AMSR2 Microwave Observations Based on Deep Learning, IEEE Transactions on Geoscience and Remote Sensing, 2025.

The main text has been amended as follows:

"Because the completion and retrieval stages are machine-learning based, specific model design choices (e.g., temporal length and incremental-learning strategy for GDCM; predictor selection, matrix size, and architectural/training techniques such as ERA5 pretraining and transfer learning for MPFNet) can affect the results. In generating DeepFlux, we adopt the peer-reviewed configurations of GDCM and MPFNet that have been evaluated through sensitivity/ablation experiments, and we summarize the key ablation evidence in the Supplement (Figs. S4–S7); full details are provided in the companion methodology papers (Wang et al., 2025; Wang et al., 2025).

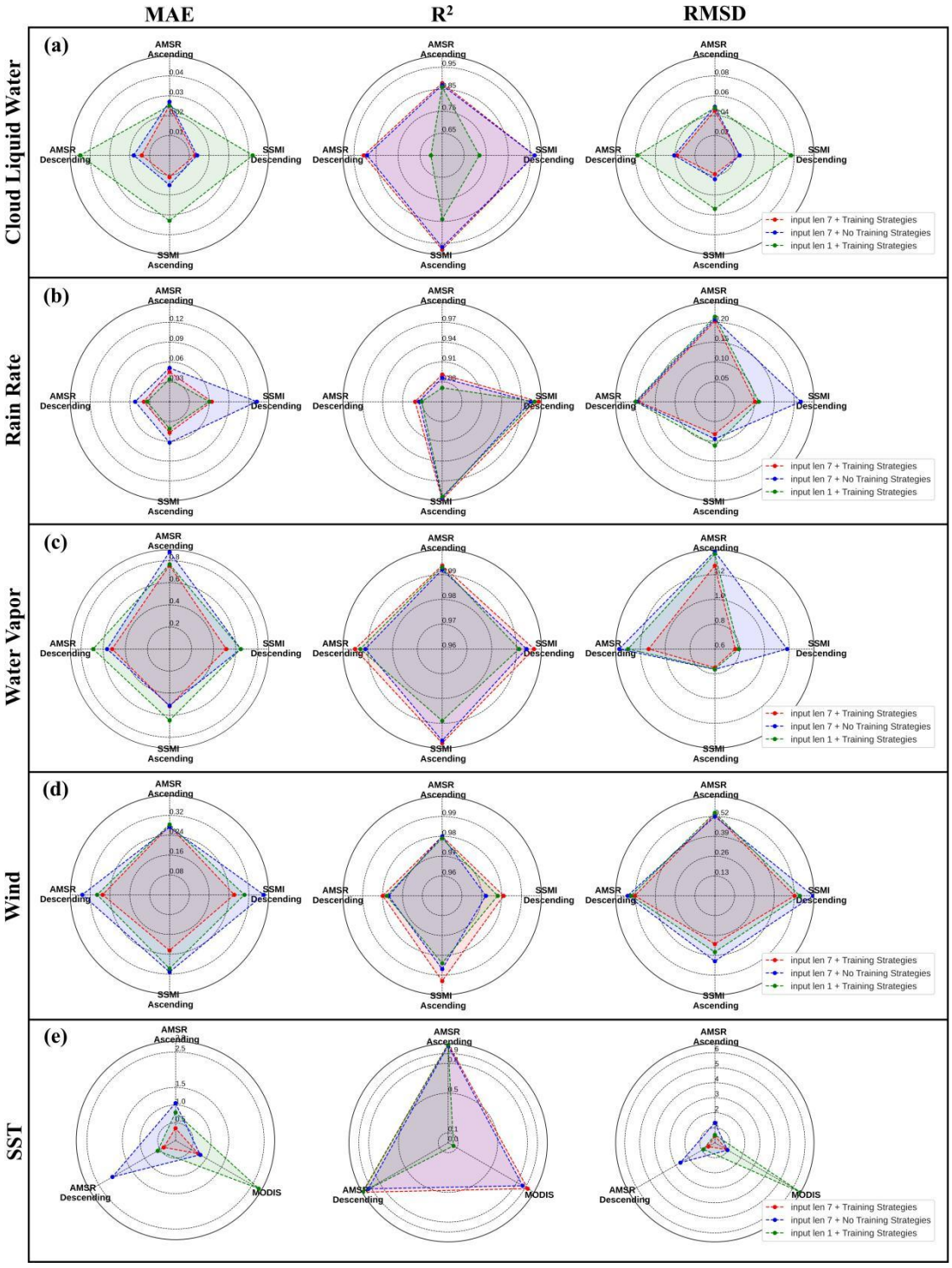


Figure *S4. Comparison of evaluation indexes of five variables of AMSR2, SSMI, and MODIS under different training strategies.*

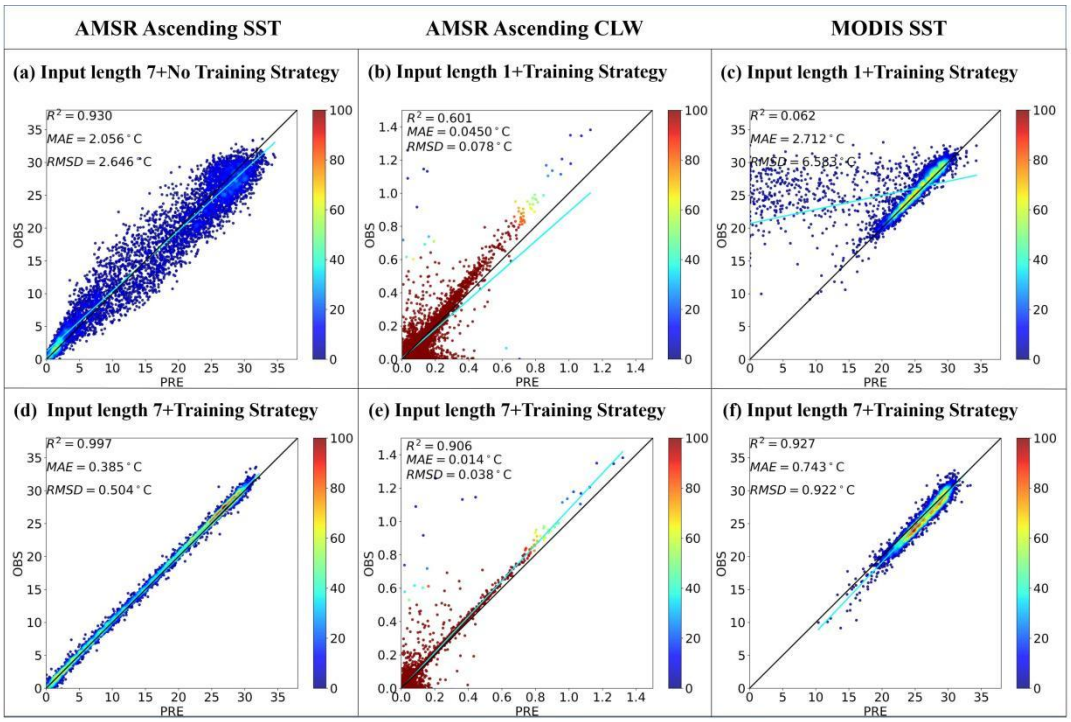
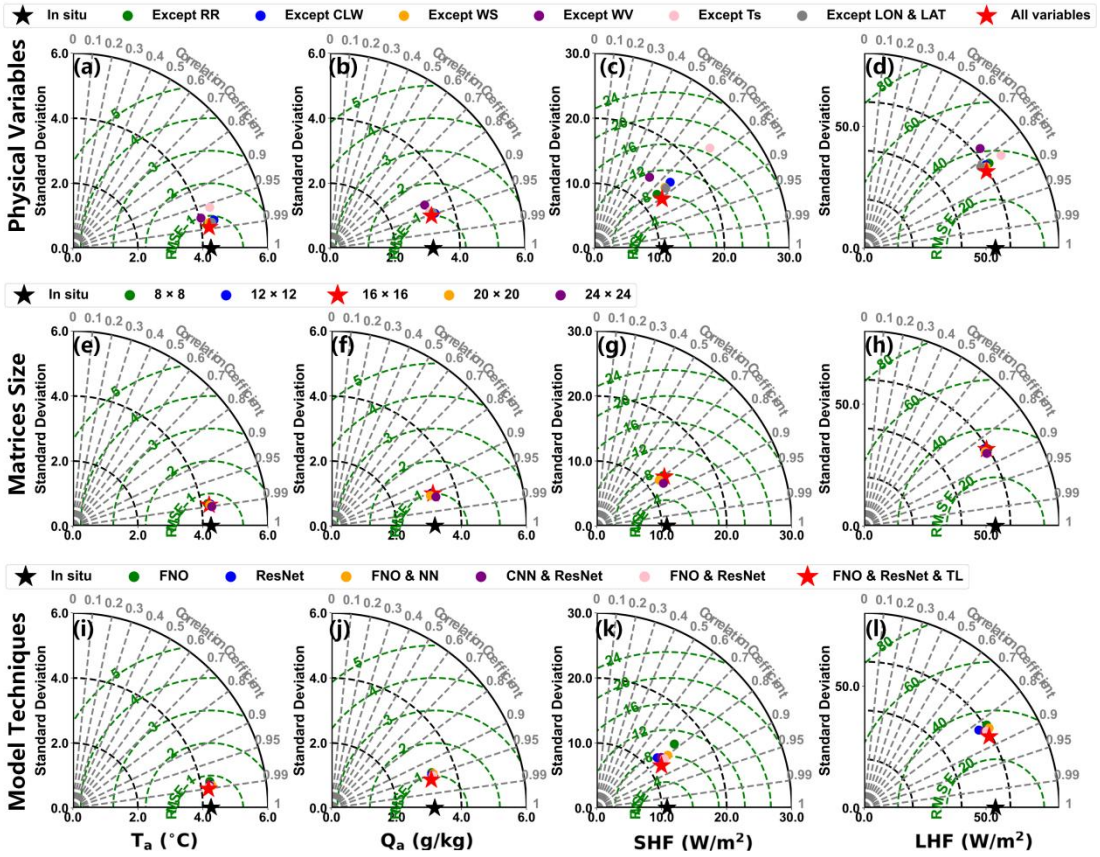


Figure S5. Scatter comparison plots of AMSR2, MODIS SST, and Cloud liquid water Descending data after GDCM model data complements.



FigureS6. Taylor Diagrams for three ablation studies on MPFNet are plotted using polar coordinates, with the radial axis representing STD and the angular axis representing CC. Green contours indicate the RMSE. Intercomparisons between in situ measurements (black pentagon) and variations in (a)–(d) variable selection, (e)–(h) matrix size, and (i)–(l) model techniques across test sets for T_a, Q_a, SHF, and LHF. The red pentagons represent the final MPFNet model configuration determined by optimal variable selection, matrix size, and model techniques.

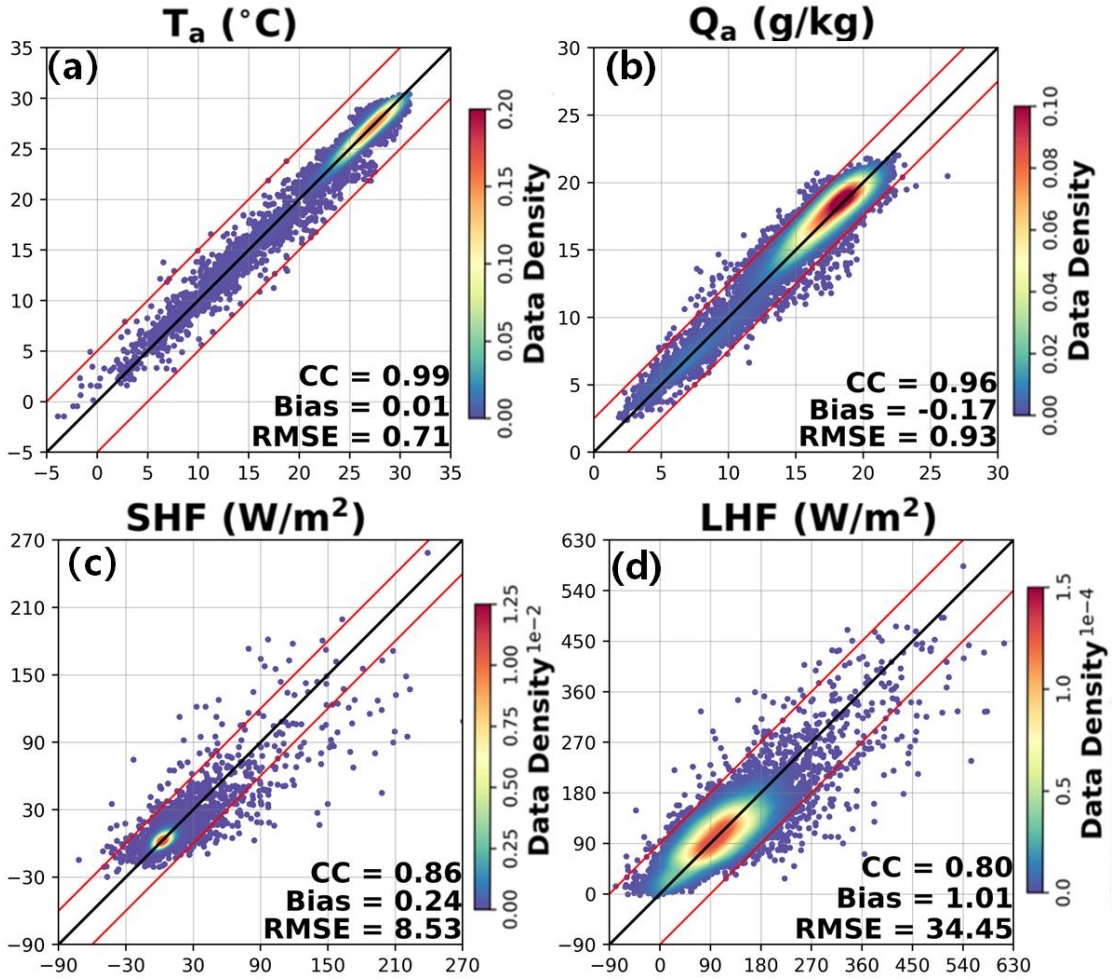


Figure S7. Scatterplots comparing retrieved values from the SSM/I-based fine-tuned MPFNet against in situ measurements for the test set, including (a) T_a, (b) Q_a, (c) SHF, and (d) LHF. The black lines in scatterplots denote the reference line with a slope of 1, indicating perfect agreement between the retrieved and observed values. The red line represents the symmetrical linear fit, indicating two STDs of the differences between the predicted and observed values, encompassing approximately 95% of the data points."

In this context, the total independence of comparison data from training data becomes crucial. But this is not always easy to be clear about, especially when re-analysis fields have been used as part of the training, as the assimilation products may well have ingested all or much of the comparison data. Comments focussed on and acknowledging any limitations of independence would help on this topic.

Response:

We clarify our split strategy and explicitly acknowledge that full statistical independence cannot be guaranteed when reanalysis fields are used for pretraining. To strengthen independence for an in-situ evaluation subset, we withhold NTAS/Stratus/WHOTS from model training and use them only for independent evaluation (monthly time series).

The main text has been amended as follows:

"In the first step, ERA5 data served as both the input and output for the pre-training phase, which adjusted the randomly initialized MPFNet to produce the pre-trained MPFNet. One thousand random points were sampled daily (at 00:00) from the ERA5-provided T_a *and* Q_a *as label data to pre-train the inversion model (MPFNet), covering the period from January 1, 1992, to December 31, 2020 (excluding 2018), with a total of 13,149,000 records. In the second step, data from SSM/I F10–F16 satellites matched with buoy observations (excluding WHOTS/Stratus/NTAS, which are reserved for independent evaluation) are used to fine-tune the MPFNet model, with 5% of the data (excluding 2018) randomly selected as the validation set for each model. Due to the earlier observation periods of F10 and F11, fewer matched records with buoy data are available, leading to overfitting during training. To address this issue, we combine F15 and F17 data—which do not overlap in time with F10 and F11—with the earlier records to mitigate overfitting during fine-tuning. In the final step, the calibration model's training set includes observed data from 1992 to 2020, excluding 2018, with 1,459,414 matched records. Data from 2018 are used as the test set, with 21,613 matched records. The detailed split of the training and test sets is shown in Table 2.*

To facilitate regional analysis and evaluation, we selected monthly averaged data from three independent buoy stations—NTAS (51 °W, 15 °N), Stratus (85 °W, 22 °S), and WHOTS (158 °W, 22.7 °N)—as additional datasets to assess the accuracy of different heat flux products under varying environmental conditions. To ensure independence, observations from these three moorings are withheld from model training and used only for independent evaluation. The time spans covered are 2002-2020, 2001-2020, and 2005-2020, respectively, as shown in Figure 4-7."

I was left a little unclear what the precise measurand heat-flux is. I infer the product aims for a global completed heat flux equivalent to the instantaneous heat fluxes one would be able to retrieve from the satellite data, and that the comparison data are matches to the nearest SSMI comparison time. If so, there is an issue with using the product for long-term change analysis in that there is a subdaily cycle in heat fluxes, and the satellite observation times are not consistent over the full period. (Targetting an explicitly daily-mean heat flux by machine learning might be a useful approach and could be validated against in situ data at high temporal resolution aggregated to daily values.)

Response:

We sincerely thank the reviewer for this insightful comment. We apologize for the lack of clarity in our original manuscript regarding the precise measurand.

To clarify, our target product is the daily-mean heat flux, not the instantaneous flux at specific satellite overpass times.

The aliasing effects are caused by sub-daily cycles (diurnal variations), combined with inconsistent, drifting satellite observation times. We agree with this assessment. It is precisely to overcome these exact limitations that our machine learning approach was designed to directly reconstruct the daily-average fields.

Our model explicitly targets the daily-mean heat flux. Furthermore, to ensure rigorous training and validation, the high-temporal-resolution *in situ* measurements (e.g., from buoys) were aggregated into daily values. By validating our machine-learning-derived daily fluxes against these *in situ* daily aggregates, we effectively bypass the sub-daily sampling biases and satellite drift issues, making the resulting dataset suitable for long-term climate analysis.

OISST is used for SST trends. This is an unfortunate choice among the available options for a long-term SST record, as OISST's operational mode of production is associated with inconsistency of bias referencing over time (Journal of Climate 34, 2923–2939 (2021)), causing relatively out-of-family trends (instability) over the period of this dataset. (See: 10.1175/JCLI-D-20-0793.1 ; https://climate.esa.int/documents/2370/SST_CCI_D5.1_CAR_v1.1-signed.pdf.)

Response:

We appreciate the reviewer's insightful comment regarding the potential temporal inconsistencies and trend instabilities introduced by OISST's operational production mode. We fully agree that long-term SST analyses can exhibit time-dependent inhomogeneities due to changes in input data streams and bias-referencing strategies, as highlighted in the literature (e.g., 10.1175/JCLI-D-20-0793.1 and the ESA SST CCI Climate Assessment Report). This is a critical consideration when using such products in climate studies.

In our study, we explicitly acknowledge that OISST is an externally blended product rather than a direct, homogeneous satellite observation that strictly matches the SSM/I record. Because it merges multiple data sources, factors such as the sparsity of input observations, empirical blending weights, and inherent spatial smoothing inevitably introduce systematic biases, often suppressing small-scale physical variations. Consequently, we do not treat OISST as an error-free ground truth. In fact, our preliminary evaluations align perfectly with the reviewer's concerns about its accuracy: as shown in Figure S3, OISST exhibits a noticeably larger Root Mean Square Error

(RMSE) relative to in situ measurements compared to the SST product used by OAFlux. This clearly indicates that the systematic errors inherent in this external SST product are non-negligible.

If left unaddressed, these inherent OISST errors, including the temporal instabilities and out-of-family trends the reviewer rightly noted, would propagate and amplify during the downstream flux-retrieval process. It is precisely to mitigate this critical issue that we designed the final step of our framework. To suppress the propagation of uncertainties stemming from the OISST inputs and to enhance the ultimate flux accuracy, we explicitly constructed the Sensible Heat Flux (SHF) and Latent Heat Flux (LHF) Calibration Models. These models are designed to rigorously calibrate the initial flux priors against high-quality in situ matchups, effectively buffering our final heat flux dataset against the systematic biases and long-term drifts introduced by the external SST input.

It is also important to clarify OISST's role within our retrieval framework. OISST is primarily used as an auxiliary predictor to provide sea-surface temperature (Ts) for the MPFNet retrieval, since SSM/I does not directly retrieve Ts. In other words, OISST is not used as a training target (“ground truth”) for MPFNet; instead, MPFNet learns the mapping from multi-source inputs to near-surface air variables (Ta, Qa) through (i) pretraining on the large-scale ERA5 dataset and (ii) transfer learning/fine-tuning using satellite–in situ matchups. This two-step procedure enables the network to adaptively correct systematic input biases when mapping to the final retrievals. When MPFNet was adapted to SSM/I F16, OISST was used as the alternative Ts input, and the accuracy reduction relative to the AMSR2-based case was minor, with RMSE increases of 16.9% (Ta), 6.45% (Qa), 23.33% (SHF), and 14.89% (LHF). We explicitly attributed this small degradation mainly to uncertainties introduced by the externally sourced OISST, while the results still demonstrate that MPFNet can be readily adapted for long-term SSM/I series applications (Figure S7). This further supports that our framework effectively contains the impact of OISST’s inherent limitations.

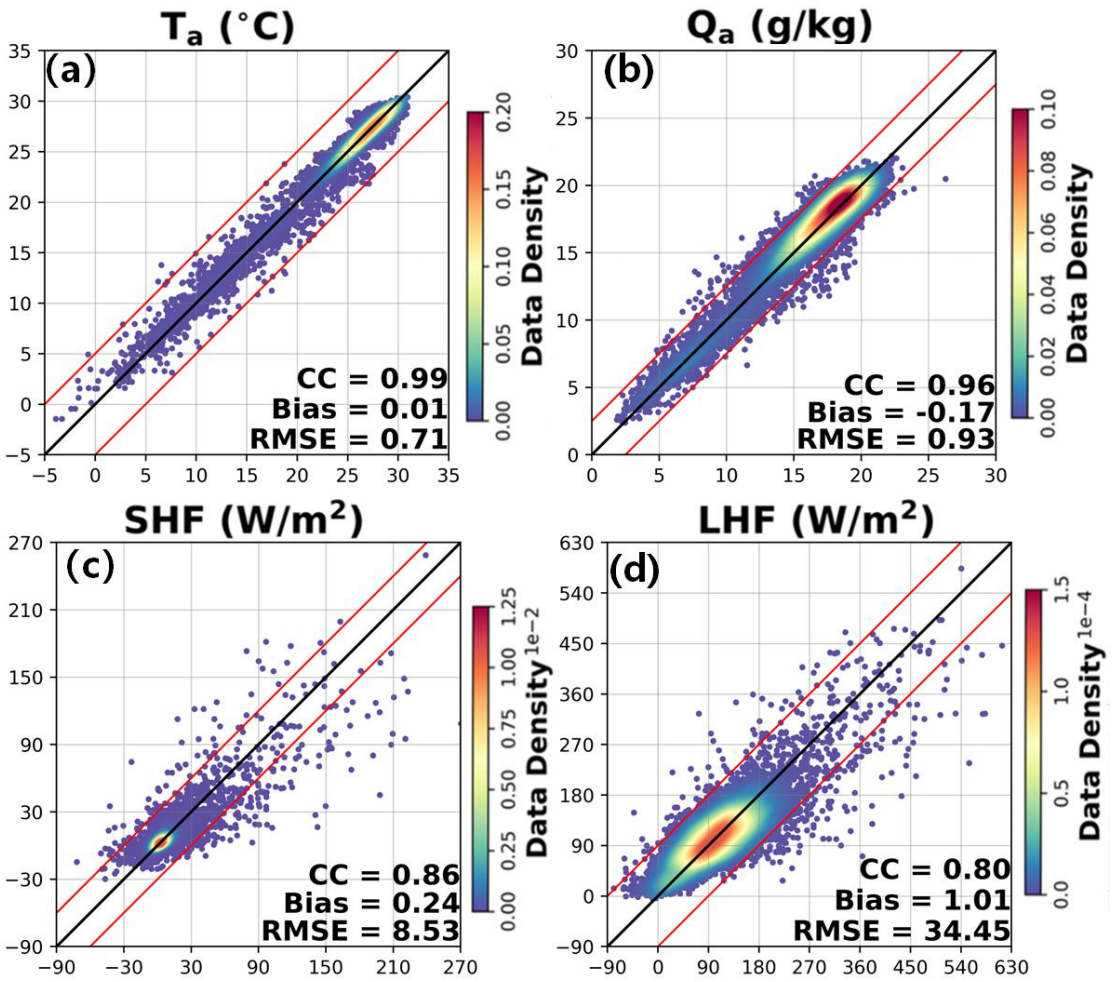


Figure S7. Scatterplots comparing retrieved values from the SSM/I-based fine-tuned MPFNet against in situ measurements for the test set, including (a) T_a, (b) Q_a, (c) SHF, and (d) LHF. The black line in scatterplots denotes the reference line with a slope of 1, indicating perfect agreement between the retrieved and observed values. The red line represents the symmetrical linear fit, indicating two STDs of the differences between the predicted and observed values, encompassing approximately 95% of the data points.

Finally, we revised the (input–output mapping and (b) the implications of potential SST-analysis inhomogeneity for trend interpretation. We also added a cautionary note that long-term trend attribution involving Ts (and derived Qs) may be affected by the choice of SST analysis product, and we direct readers to treat the SST-related trend interpretation accordingly.

The main text has been amended as follows:

"The NOAA OISST dataset integrates observations from multiple platforms, including satellite infrared and microwave sensors, ship measurements, and buoy data. Using an optimal interpolation algorithm, it fills spatial gaps and merges data to produce daily global SST fields at a spatial resolution of 0.25°×0.25 °, covering the period from September 1981 to the present. In this study, we use global SST data from

OISST v2.1 (January 1, 1992 to December 31, 2020) along with GDCM-completed SSM/I SSW, CLW, WV, and RR data over the same period as input for the MPFNet model to retrieve global and . Note that OISST is used as an auxiliary input to provide Ts because SSM/I does not directly retrieve Ts. Therefore, OISST is not used as a training target ("ground truth") for MPFNet. MPFNet is first pretrained on the large-scale ERA5 dataset and then fine-tuned using satellite–in situ matchups, enabling adaptive adjustment of the input–output mapping and mitigating moderate systematic biases in externally sourced Ts. This design is consistent with our published MPFNet transfer experiments, in which using OISST as an alternative Ts input for SSM/I resulted in only minor performance degradation and still demonstrated robust adaptability for long-term SSM/I-series applications (Wang et al., 2025).

Once the input data fields are complete, the retrieval process commences using the MPFNet architecture (Wang et al., 2025). The first step is the primary retrieval of T_a *and* Q_a*. To address the sample imbalance inherent in the matched satellite-in situ dataset, the MPFNet model is first pretrained on ERA5 data, then fine-tuned using a training set constructed from matched remote sensing and in situ observations. Because OISST provides Ts as an external predictor, we rely on the ERA5 pretraining + matchup fine-tuning strategy to adapt the input–output mapping and reduce sensitivity to moderate systematic biases in the Ts input, as demonstrated in our published MPFNet transfer experiments (Wang et al., 2025). This process yields initial global and fields, from which preliminary LHF and SHF are calculated using the bulk aerodynamic formulas (Equations 1 and 2).*

In the equatorial central and eastern Pacific regions, the positive trend of Q_a *weakens and even shows a slight negative trend in some local areas, with significant spatial variability. We caution that inferred long-term trends in* T_s *(and derived* Q_s*) may depend on the chosen SST analysis, because operational changes and time-dependent bias referencing in some SST analyses can introduce inhomogeneities that affect trend estimates (e.g., Yang et al., 2021; ESA SST CCI Climate Assessment Report). Therefore, the SST-related trend interpretation should be viewed as supportive evidence, while the primary contribution of DeepFlux lies in the observation-constrained orbit-sampled* T_a*,* Q_a*, SHF and LHF fields validated against in situ measurements."*

I would like to see in the paper a comparison of the accuracy statistics for matches that were present in the SSMI swaths compared to the infilled times-and-places. This would be a good measure of the effectiveness of the infilling in providing a "daily" complete product.

Response:

We agree and thank the Reviewer for this helpful suggestion. DeepFlux is a dataset paper, so in the main text, we intentionally keep method/diagnostic figures concise and refer readers to the companion methodology papers for detailed performance

evaluations. In particular, the completion step is based on our Generalized Data Completion Model (GDCM), which explicitly separates data-missing and non-missing regions and reports quantitative error distributions and bias statistics for both subsets (e.g., error probability distributions and bias tables for missing vs. non-missing regions).

To address the Reviewer's request within this ESSD paper, we add the key completion-performance evidence to the Supplement, including (i) missing vs non-missing diagnostics from the GDCM evaluation and (ii) (for completeness) the MPFNet robustness/ablation evidence that supports the stability of the downstream retrieval stage. We also add a short statement in the main text pointing to these Supplementary figures/tables, so readers can directly assess the effectiveness of the infilling step in providing a daily spatially complete product.

The main text has been amended as follows:

"Figure 2 presents Taylor diagrams and Figure 3 presents scatter plots comparing various heat flux products with in situ observations, while Table 4 summarizes their performance in terms of RMSE and CC. Among all datasets evaluated, the SSM/I heat flux product shows the highest accuracy and consistency with in situ data, achieving the lowest RMSE and the highest CC for T_a, Q_a, SHF, and LHF(Figure 2).

To specifically assess the effectiveness of the GDCM infilling, we provide supplementary diagnostics that compare errors in data-missing versus non-missing regions (see Supplementary Fig. S4-5 and Table S1-S2), based on the completion-validation framework reported in the companion GDCM paper (Wang et al., 2025a)."

Wang, H., Zhou, Y., and Li, X.: GDCM: Generalized data completion model for satellite observations, Remote Sensing of Environment, 324, 114760, 2025.

Overall, the paper is well written and presented. There is inconsistency in acronyms being presented within and without being italicised, and sometimes named differently in figures (e.g. SSW and WS). Table 1 is very confusing and needs to be aligned in a way the reader can understand what is connected to what.

Response:

We reformatted Table 1 with gridlines and clarified inputs/algorithms/periods. We will also ensure acronym consistency across text and figures (e.g., SSW vs WS) in the final revised figures.

Table 1: Table of characteristics of different heat flux products

Input Data	*Algorithm*	*Heat Fluxes Product*	*Spatial resolution*	*Temporal resolutio*	*Period of availability*	*Source*

				n		
SSM/I to SSMIS (RSS V7) + OISST; ERA5 only for missing days	GDCM + MPFNet COARE 3.6	DeepFlux	1°× 1°	Daily	1992.01.01–2020.12.31	IOCAS
	MLP COARE 3.0	OHF CDR	0.25° × 0.25°	3-hourly	1988.01.01-2021.08.31	NOAA
Reanalysis	ECMWF Scheme	ERA5	0.25° × 0.25°	Hourly	1940.01.01-Present	ECMWF
	NCEP Scheme	NCEP	T62 Gaussian	6-hourly	1948.01.01-Present	NOAA
Blended	Regression COARE 3.0	IFREMER v4.1	0.25° × 0.25°	Daily	1992.01.01-2018.12.31	IFREMER
	Least Squares COARE 3.0	OAFlux	1° × 1°	Daily	1981.01.01-2022.12.31	WHOI