Thanks to all the reviewers for their thoughtful comments and suggestions, which have strengthened the manuscript. We have engaged with these suggestions and have responded below in blue, as can be seen in the new version of this manuscript.

This manuscript presents the Coral Reef In Situ Product (CRISP), a global compilation of in situ temperature records from coral reef environments, comprising 2,548 logger deployments and more than 595 million observations. The effort involved in assembling, standardising and making these disparate records accessible is substantial, and the resulting archive has clear potential value for studies of reef thermal environments, coral bleaching, model validation and broader comparative analyses.

However, in its present form I do not think the manuscript provides sufficient information to assess the quality, comparability and appropriate use of the compiled records. This is particularly important given the considerable heterogeneity among the underlying datasets in sampling interval, deployment depth, logger model, deployment configuration and metadata completeness. The manuscript currently places considerable emphasis on the motivation for compiling in situ temperature observations, but substantially less on demonstrating how the resulting data product has been quality controlled and how users should account for differences among records. The description of quality control is especially limited, relying heavily on visual assessment without sufficiently explicit or reproducible criteria, while the procedure used to infer and correct timezone offsets also requires much more justification.

I therefore think CRISP could become a useful community resource, but the manuscript requires substantial revision before users can confidently interpret and combine records across the archive. In particular, the authors should provide a much more detailed and reproducible description of quality-control procedures, quantitatively characterise the major sources of heterogeneity within the dataset, and provide clearer guidance on the suitability of different records for resolving thermal variability across different temporal and spatial scales.

The quality-control procedures require substantially greater documentation and, ideally, quantitative characterization. CRISP combines a very large number of records collected using different instruments, deployment approaches, sampling intervals and metadata standards, yet the present description of QC relies heavily on subjective visual assessment. For a data product intended for broad reuse, users need to be able to understand precisely how observations were assessed, what criteria were used to flag or modify records, how ambiguous cases were treated, and how many observations or deployments were affected. This is particularly important for the treatment of anomalous values and inferred timezone corrections. I think a reproducible QC protocol, accompanied by a clear description of QC flags and summary statistics on their application, is essential.

More generally, the manuscript needs to engage much more directly with the substantial heterogeneity among records and what this means for their comparability and fitness for different analyses. Sampling intervals span several orders of magnitude, deployment depths vary greatly, and logger models and deployment configurations differ in ways that can materially affect the temperatures recorded. These are not simply ancillary metadata differences: they determine whether a record can resolve diel variability, short-lived extremes, internal-wave events or other processes that motivate the development of CRISP in the first place. The manuscript should therefore provide a much more quantitative characterization of these sources of heterogeneity and clearer guidance on how users should filter or stratify the dataset for

different applications. Without this, there is a risk that the apparent standardization of the archive implies a degree of comparability among records that the underlying observations do not necessarily support.

Thank you for the positive feedback and constructive criticism. We agree with many of these sentiments and have provided a more in-depth description of quality control and literature review procedures as a result. Additionally, we have made greater efforts to stress the heterogeneity of the data product to those who may be unfamiliar with the diverse nature of logger data. See comments below.

L11: The word "still" implies that satellite SST products might ultimately be expected to resolve in situ reef temperatures at these spatial and temporal scales. I do not think that implication is warranted; these are fundamentally different measurements with different spatial and temporal representations.

Thank you, the word "still" was removed from Line 11.

L37: I do not think it is accurate to state that NOAA Coral Reef Watch "monitors bleaching events globally". CRW monitors remotely sensed thermal conditions associated with bleaching risk; it does not directly observe bleaching events. The wording should be revised accordingly.

Line 37 was adjusted accordingly to reflect risk assessment rather than direct bleaching event monitoring.

Additionally, lines 37-45 were edited in response to a comment by Reviewer #1. We believe these edits reflect a change in language that addresses the comment here as well.

L42: This discussion should also acknowledge the well-established limitations of DHW as a predictor of observed bleaching severity. DHW captures only part of the spatial and temporal variability in bleaching response, particularly where local hydrodynamics and subsurface thermal structure strongly modify reef temperatures. Relevant literature demonstrating these limitations should be cited here.

The manuscript now includes substantial discussion regarding the reviewer's comment about DHW limitations, with relevant literature cited. See edits below:

Lines 50-54 edited to "However, satellite-based SST products can only provide daily observations at resolutions 1 km or coarser—a scale too broad to capture the spatial patchiness of reef habitats. Consequently, metrics like satellite-based DHW lack the spatial and temporal resolution required to resolve fine-scale heat stress variability across individual reef zones and are limited by processes such as mesoscale eddies (Wyatt et al., 2023) and internal waves (Storlazzi et al., 2020) that modify the water column temperature structure."

Additionally, we have addressed the limitations of satellite-based heat stress metrics in the following paragraph, see Lines 67-73.

L53-55: This passage warrants referencing.

Edits to the above section regarding the previous comment rendered these lines repetitive so they were rewritten and referenced accordingly. See edits below:

Lines 56-57 edited to "By failing to capture fine-scale, within-reef thermal variability, satellite-based monitoring can miss critical heat stress events that can be resolved by *in situ* loggers (Lippert et al., 2026)."

L62-63: The statement that discrepancies "can exceed 0.5 °C" substantially understates the magnitude of the differences documented in some reef environments. Several studies report departures from satellite SST of several degrees during processes such as upwelling and internal-wave forcing. The wording should better reflect the range of observed discrepancies.

We have included additional literature (Wyatt et al., 2023 and Margaritis et al., 2025) to broaden the language used in addition to providing deeper context to the range of discrepancies observed. See edits below:

Lines 63-67 were edited to "Multiple studies have reported strong correlations between satellite and *in situ* temperatures at broad scales, but found systematic differences exceeding 0.5 °C and localized discrepancies of up to 5 °C during oceanographic processes like upwelling, internal waves, and sub-diurnal anomalies (Castillo and Lima, 2010; Claar et al., 2019; DeCarlo and Harrison, 2019; DeCarlo et al., 2021; Margaritis et al., 2025; Schramek et al., 2019; Wyatt et al., 2023)."

L67: Again, an important limitation is that, regardless of the underlying data source, DHW is a temporally coarse metric, integrating thermal stress over a 12-week (~3-month) period. It is therefore unlikely to adequately capture short-duration heatwave events and other high-frequency thermal variability relevant to coral bleaching.

Short-term heat stress was added as a limitation (Lines 70-71), reflecting previous commentary throughout the manuscript. DeCarlo et al., 2017 was cited as well.

L102: What was the rationale for using Rich et al. (2024) as the basis for identifying the 314 publications screened for data? That study does not appear to constitute a comprehensive census of the reef thermal-environment literature. Please explain how the literature search and selection criteria were defined and what important datasets or study classes might consequently have been omitted.

It is our understanding that Rich et al., 2024 intended to represent all known literature using temperature loggers deployed on coral reefs and published between 2013-2022. It was, therefore, a great starting point for literature that may contain temperature logger data. Additionally, we have done our own searching for time periods not covered using Google Scholar and searching for files utilizing in situ temperature logger data on coral reefs. In some cases, contacting authors led to additional datasets that were unpublished. There was also an opportunity to present CRISP at ICRS in July 2026, which led to some additional dataset contributions. We argue we have likely captured the vast majority of data, though it is quite possible we have missed some datasets. With the publication of CRISP, we hope more data will be submitted—including unpublished datasets—and can be included in future versions of CRISP.

Additional criteria were included for how the contact study was performed to account for individual authors based on Rich et al., 2024. See edits below:

Lines 103-110 were edited to: "To acquire *in situ* temperature logger data in published literature, an additional 323 peer-reviewed papers (based on Rich et al., 2024) that mention the use of temperature loggers

were reviewed for data compilation. The initial list of 397 entries in Rich et al. (2024) was reduced to 323 unique publications after accounting for redundant listings that employed multiple logger types. Temperature records were then extracted from journal articles, supplementary materials, and associated repositories when available. For time periods not covered in Rich et al., (2024), we performed additional searching for files utilizing *in situ* temperature logger data on coral reefs. While it is possible some available datasets were not yet incorporated in CRISP, our approach likely captures the vast majority of data available through December 2025."

Lines 116-117 were edited to: "To directly assess data recovery rates for missing or inaccessible datasets, we consolidated papers by lead author to avoid duplicate outreach."

L173–177: The QC procedure requires substantially more detail for the resulting data product to be reproducible and its reliability assessable. "Careful visual observation" is inherently subjective. Please specify the criteria used to identify anomalous observations, whether thresholds or rules were applied, how deployment/retrieval artefacts were distinguished from genuine environmental extremes, who performed the assessment, whether classifications were independently checked, and how many observations or records were ultimately flagged. Equivalent detail is also needed for the timezone QC procedure described below.

Thank you for the comment and the chance to clarify that we are not deleting any data but rather providing flags. Users may choose to use those flags or not, it is at their own discretion. In developing our QC workflow, we initially tested automated standard deviation thresholding, but found it missed certain erroneous records while overly flagging valid extreme events. While subjective, visual screening was necessary to identify and flag aggressive anomalies. We have added an additional line to reflect this sentiment in the manuscript:

Lines 200-204 were added, "Visual flagging across both flag vectors were reviewed by two authors to ensure consistency, providing users with a baseline quality assessment. To avoid introducing unfounded assumptions, ambiguous cases were neither flagged nor adjusted, keeping the original data fully intact. Since raw data remain intact, users retain the choice to utilize, adjust our flags, or even establish new flags, entirely at their discretion."

L177: "No raw data were deleted" appears inconsistent with the preceding statement that measurements "were removed". Presumably the observations were retained but flagged/excluded from downstream use; please clarify the terminology.

Thank you for the chance to clarify. The terminology was edited to reflect the reviewer's request. See below:

Line 185-190 were edited to: "These data measurements were flagged utilizing a cleaning system that allowed team members to identify anomalous and erroneous measurements within a given temperature time series through careful visual observation (Figure 3). No raw data were deleted; rather, a Quality Control (QC) Flag vector of the same length as the temperature record was created to track flagged

observations. As a result, temperature measurements flagged during the pre-processing stage remain in the published dataset but can easily be masked or excluded from analysis using the QC Flag vector."

L180–184: Inferring timezone offsets from the phase of the diel temperature cycle seems potentially problematic in environments where temperature variability is driven primarily by tides, internal waves, upwelling or advection rather than local solar heating. Please provide the decision criteria used for these corrections, indicate how ambiguous cases were handled, and ideally retain a flag distinguishing reported timestamps from inferred/corrected timestamps.

Similarly to the above comment on QC Flags, visual assessment was the most viable option for determining misalignments and users have full flexibility to disregard flags. We have included a short description (see edits below) with a figure displaying the process by which we assessed the time zones and have retained an additional flag "UTC Flag" in CRISP to be published with the next version (CRISPv2.nc).

Lines 192-200: "Time zone quality control was completed by visual assessment of diurnal temperature cycles for records lacking explicit time zone metadata. Diurnal temperature anomaly cycles corrected for seasonality were plotted in each logger's geographical time zone and in UTC (Figure 4). Occasionally, misalignments between diurnal cycles and time of day were used to infer and correct time zone offsets for loggers in which the time zone of the data as recorded was not disclosed. For example, logger record L02148 originally lacked time zone metadata; when assuming the data were stored in local time, peak diurnal temperature occurred between 00:00-02:00. Conversely, if the data were assumed to be in UTC and then converted to local time, the peak diurnal temperature occurred in the expected midday window. This logger was determined to have been recorded in UTC and was adjusted accordingly in CRISP. A UTC Flag has been retained for users to identify which time zones have been modified by the authors."

L205–207: The sampling resolution should also be stated here (i.e., in the figure description/caption). Temporal resolution is perhaps one of the most important factors determining whether an in situ record is meaningful for resolving the thermal variability of interest.

We made edits to satisfy the reviewer's request, see new figure caption below:

Lines 228-230 "Figure 5. Example temperature time series extracted from the CRISP dataset. Logger L00404 was deployed at Crocker Reef, Florida Keys, United States of America over a span of 9.2 years at a sampling interval of 15 minutes. The logger recorded temperatures ranging from 19.7°C to 32.4°C."

L243-244: This discussion needs greater nuance. Solar heating of an exposed logger may represent a measurement bias relative to bulk water temperature, but corals themselves can also experience substantial radiative heating. The manuscript should therefore be explicit about what environmental quantity CRISP is intended to represent (e.g., bulk water temperature versus the thermal environment experienced by benthic organisms), and about the extent to which shading or sensor-specific radiative responses affect that interpretation. More broadly, I am not entirely convinced by the framing in Rich et al. (2024), where exposure to solar radiation is treated principally as a source of logger bias; in shallow reef environments, eliminating radiative heating from the measurement does not necessarily provide a better representation of the thermal conditions actually experienced by corals.

We agree with this sentiment that there is arguably a distinction between measured water temperature and stress felt by corals. CRISP is intended for use by researchers interested in thermal stress, we are not attempting to resolve the compounded effect of light/radiative heating and thermal stress. Additionally, while the notion that loggers could capture some form of radiative heating imposed on sunlit surfaces, we argue this would need more experimental testing. For instance, the plastic casing and internal airspaces of most loggers likely respond differently to radiation than a coral surface. Therefore, CRISP should be treated as the bulk water temperature. See addition below:

Lines 274-276 "However, is important to note that CRISP is intended to capture water temperature on reefs rather than quantify the combined temperature and light stress experiences by corals, and data should be treated as such regardless of shading or model type."

L239-261: I think that, as noted above, sampling interval is perhaps one of the most important limitations when interpreting these data, together with logger model and deployment depth. Sampling intervals in CRISP span several orders of magnitude, meaning that records differ fundamentally in their ability to resolve diel variability, short-lived extremes and high-frequency oceanographic processes. Likewise, some logger models and deployment configurations may yield substantially less robust measurements than others. These issues warrant considerably greater treatment in the limitations section, including their implications for comparing or combining records across the dataset.

We agree, sampling frequency is certainly a limitation between all these loggers. Some intervals will only be able to resolve certain phenomenon. This is reflected in the limitations section now.

Lines 284-287 were added, "As a comprehensive collection of *in situ* logger records, CRISP encompasses diverse spatial and temporal scales. Variations in sampling interval result in varying capability to capture and resolve thermal events between records. Subsequent analysis of loggers should consider differences in sampling resolution between loggers and how this might influence the user's application."