Response to Comments by Anonymous Referee #2:

Review Summary

This manuscript describes a new dataset calculating river ice concentration and ice phenology across major Arctic rivers using 500 m resolution MODIS data. The paper is clearly written, and the dataset is clearly structured and includes relevant information needed for others to use the data. I think the paper could benefit from having a stronger justification about why this particular method is better than other existing methods (and what sorts of analyses you could do with this dataset, that you couldn't do with existing Landsat-based methods). I also have some concerns, described below, about how the water masks were developed, particularly in sub-MODIS pixel width rivers, as well as with the use of single reflectance or NDSI thresholds when there are so many mixed land/water pixels.

Response:

Thank you very much for your positive overall assessment of our manuscript and dataset, as well as for your constructive comments.

We have tried our best to carefully address the concerns you raised. In particular, we have strengthened the manuscript to better clarify the specific value of our MODIS-based approach relative to existing datasets. In this work, we relied solely on actual daily MODIS observations to directly characterize river ice conditions and derive river ice phenology at high temporal resolution. This allows us to provide temporally continuous river ice information and to avoid potential modelling uncertainties that may arise when lower-temporal-resolution Landsat observations are used to infer phenological metrics.

We also appreciate your important concerns regarding the development and applicability of the river masks, especially for reaches narrower than a MODIS pixel. Taking these concerns into account, together with the related comments from Anonymous Referee #1, we have conducted additional supplementary analyses to better assess and clarify this limitation. Corresponding revisions have been added to the manuscript.

In addition, we now explicitly discuss the potential influence of mixed pixels in our threshold-based classification framework. While the combined thresholds method used in this study has previously been demonstrated to be robust for large-scale snow/ice mapping, we acknowledge

that mixed land–water pixels may still affect retrieval accuracy in narrower river sections. We have therefore revised the Data uncertainty to more clearly explain this issue, while also noting that such mixed-pixel effects are mainly limited to a relatively small proportion (overall 12.0%) of narrow reaches, especially in smaller tributaries.

We sincerely thank you again for these valuable suggestions and hope our revisions address your concerns.

Major comments

I have three major comments regarding the paper.

- My understanding of how water masks were generated is that first, centerlines associated with rivers wider than 250 m were extracted from GRWL. The centerlines were then buffered by 250 m (to generate 500 m-wide river polygons), and then NDSI was evaluated within this mask for all rivers, regardless of original river width. There are several potential challenges that I see from this approach. In the case of river reaches whose 'average' width is close to 250 m, there are likely sections that are narrower than 250 m. For those reaches, and reaches closer in width to 500 m, the 500 m spatial resolution MODIS pixels observing the buffered area (used for the water mask) are likely to observe a lot of land, rather than water. It is thus possible that the change in NDSI being observed is snowmelt on land rather than ice breakup. While one could argue that snowmelt and ice breakup should be correlated, this is not always the case. For example, in braided rivers, sometimes snow melts off the land and sandbars prior to ice breakup – or sometimes one small channel breaks up but the rest of the braids are still frozen and appear white (or didn't have water in them during freeze-up). Point being, breakup detection is complicated in narrow rivers, particularly if any of them are braided! I would be curious to know what fraction of the studied river reaches are in the range of 250-500 m wide, and how overall accuracy (RIC, breakup date, or freeze-up date) in these reaches compares to overall accuracy in the wider reaches. On the other side of things, there are rivers whose width is much wider than 500m, and those areas are thus potentially being excluded by the 500 m buffer water mask being used. If there are large

islands in these wider rivers, the water mask could be picking up land rather than water. The work may be improved by a more detailed water mask.

Response:

We thank the reviewer for this thoughtful and highly constructive comment. We agree that the suitability of the water mask is critical for interpreting MODIS-based river ice dynamics, particularly in relatively narrow or morphologically complex reaches. We are sorry that our previous description of the river mask construction was not sufficiently clear and may have led to misunderstanding.

First, we would like to clarify that the buffered river mask was not generated by buffering the river centerline (polyline) from the Simplified GRWL Vector Product. Instead, the river corridor mask was derived from the spatially referenced GRWL Mask raster, which was converted to polygon features and then expanded with an additional buffer to better accommodate seasonal expansion and contraction of the active channel. By contrast, the purpose of introducing the Simplified GRWL Vector Product was only to screen out reaches that are too narrow for reliable analysis with MODIS, by retaining river segments with a mapped maximum width (width_max) greater than 250 m. We apologize for the misleading wording in the original manuscript and have now revised the description accordingly in **Sect 2.2.1 Spatial reference data**.

Regarding the reviewer's concern about narrower reaches (actual width in the range of 250-500 m wide), we fully acknowledge that these narrower reaches are more challenging for breakup detection. In such reaches, especially where braided morphology is present, MODIS pixels may include substantial fractions of adjacent land or exposed bars, and the observed NDSI signal may partly reflect terrestrial snowmelt rather than purely river ice decay. To address this concern, we have now conducted an additional quantitative assessment of the prevalence and performance of these narrower reaches.

Specifically, we classified the retained river network into narrower reaches and wider reaches (> 500 m), and summarized their proportions in the revised manuscript and **Supplement** (Table S3). Across the six major rivers, the length-weighted proportion of narrower reaches is 12.0%, with the Mackenzie River showing the lowest proportion (7.2%). This indicates that the majority of the analyzed river network corresponds to reaches whose widths are broadly

comparable to, or exceed, the effective MODIS mapping scale, whereas narrower reaches represent only a limited fraction of the study domain.

We further re-evaluated the retrieval accuracy separately for narrower and wider reaches. For river ice concentration (RIC), comparison with Landsat- and Sentinel-2-derived reference maps (Figs. I and II) shows that the point-number-weighted mean overall accuracy across the six rivers is 0.81 for narrower reaches and 0.80 for wider reaches, both of which are essentially comparable to the value for the full river network (0.80). These results suggest that, although narrow reaches are theoretically more vulnerable to mixed-pixel effects, their aggregate RIC accuracy remains comparable to that of wider reaches.

We also reassessed river ice phenology performance for the two subsets. The validation results (Figs. III-V) indicate that the error level for freeze-up date (FUD) is broadly similar among narrower, wider, and all reaches, with mean absolute errors (MAEs) of roughly 11 days. A similar pattern is found for breakup date (BUD) when validated against the "start of ice drift", with MAEs ranging from 8.0 to 8.8 days. When using the alternative breakup reference "End of ice", the contrast becomes somewhat larger, but the error for narrower reaches (10.2 days) still remains acceptable and is even slightly lower than that of the full-network result (11.4 days), whereas the wider reaches yield an MAE of 12.9 days. Overall, these additional analyses indicate that the retained narrower reaches do not introduce disproportionate uncertainty at the network scale, although we agree that they remain intrinsically more challenging in individual cases.

Concerning your point about much wider rivers, we would like to clarify that such reaches were not truncated to a fixed 500 m river width. Rather, the buffer was applied outward from the river corridor boundaries to allow for seasonal variability in wetted extent. We have revised the manuscript to make this buffer strategy explicit in **Sect 2.2.1 Spatial reference data**.

With respect to islands and bars within wide river channels, the GRWL-based river mask already distinguishes the actual watercourse network—including main channels and anabranching channels—from non-water geomorphic units such as larger vegetated islands and sedimental mid-channel/braided, point, or mouth bars. Consequently, large islands are generally excluded from the base river corridor mask. We agree that the additional shoreline buffer could still introduce limited overlap with the margins of some islands or bars. However, for larger

islands, the effect is expected to be minor because their characteristic dimensions are typically much greater than the applied buffer width (0.5 km), so only a relatively small marginal fraction is affected. For smaller exposed bars, some overlap may occur, but their areal contribution is very small relative to the total river corridor area across the six major basins. We have therefore added text to **Sect 6 Limitations and uncertainties** to acknowledge this source of uncertainty while noting that its influence on basin-scale results is expected to be limited.

In summary, we appreciate the reviewer's insightful suggestion. In response, we have now:

(1) clarified the actual procedure used to construct the river corridor mask;

(2) quantified the proportion of retained narrower reaches;

(3) conducted separate accuracy assessments for narrower and wider reaches for both RIC and phenology; and

(4) revised the discussion of uncertainties associated with braided sections, exposed bars, and large islands in wide channels.

We believe these additions substantially improve the transparency of the methodology and better define the scope and limitations of the dataset.

Table S3: Summary of the river centerline length for the six major Arctic rivers in this study.

	Analyzed river segments length (km)	Narrower reaches length (km)	Ratio
Mackenzie	11,890.13	858.79	7.2%
Yukon	20,919.17	2,326.00	11.1%
Kolyma	19,152.67	2,081.93	10.9%
Lena	62,037.11	8,377.70	13.5%
Yenisey	57,836.47	6,377.58	11.0%
Ob	35,586.66	4,805.59	13.5%

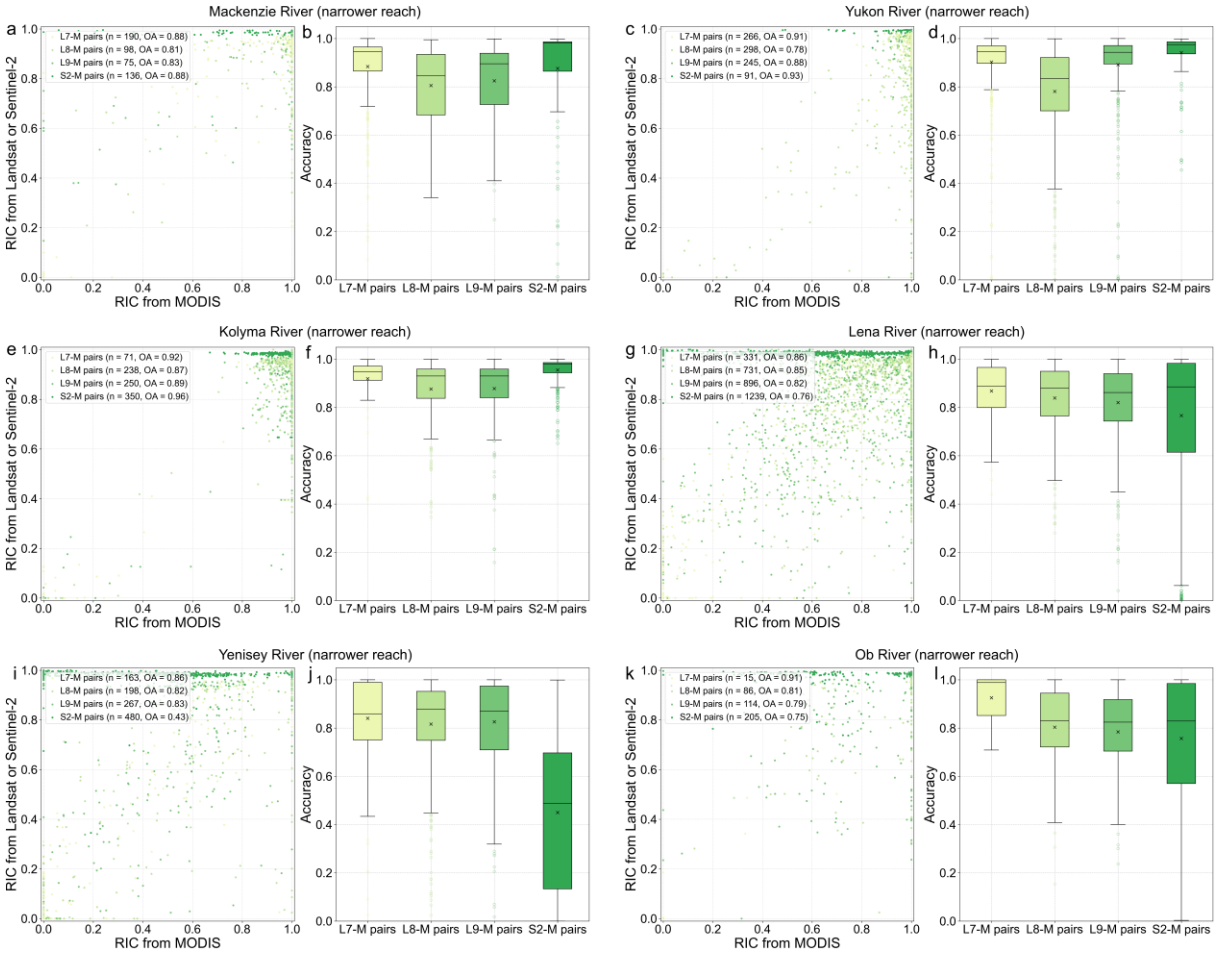

Figure I: Accuracy assessment of MODIS-derived RIC against high-resolution reference data from Landsat and Sentinel-2 across the narrower reaches of six major Arctic rivers. Left panels show pixel-wise scatterplots comparing MODIS-based RIC with higher-resolution estimates from Landsat 7, Landsat 8, Landsat 9, and Sentinel-2. Right panels present the corresponding validation accuracies for each sensor pairing (L7-M, L8-M, L9-M, and S2-M). In the boxplots, the central line indicates the median, and the cross symbol denotes the mean value. Subplots a, b correspond to the Mackenzie River, c, d to the Yukon River, e, f to the Kolyma River, g, h to the Lena River, i, j to the Yenisey River, and k, l to the Ob River.

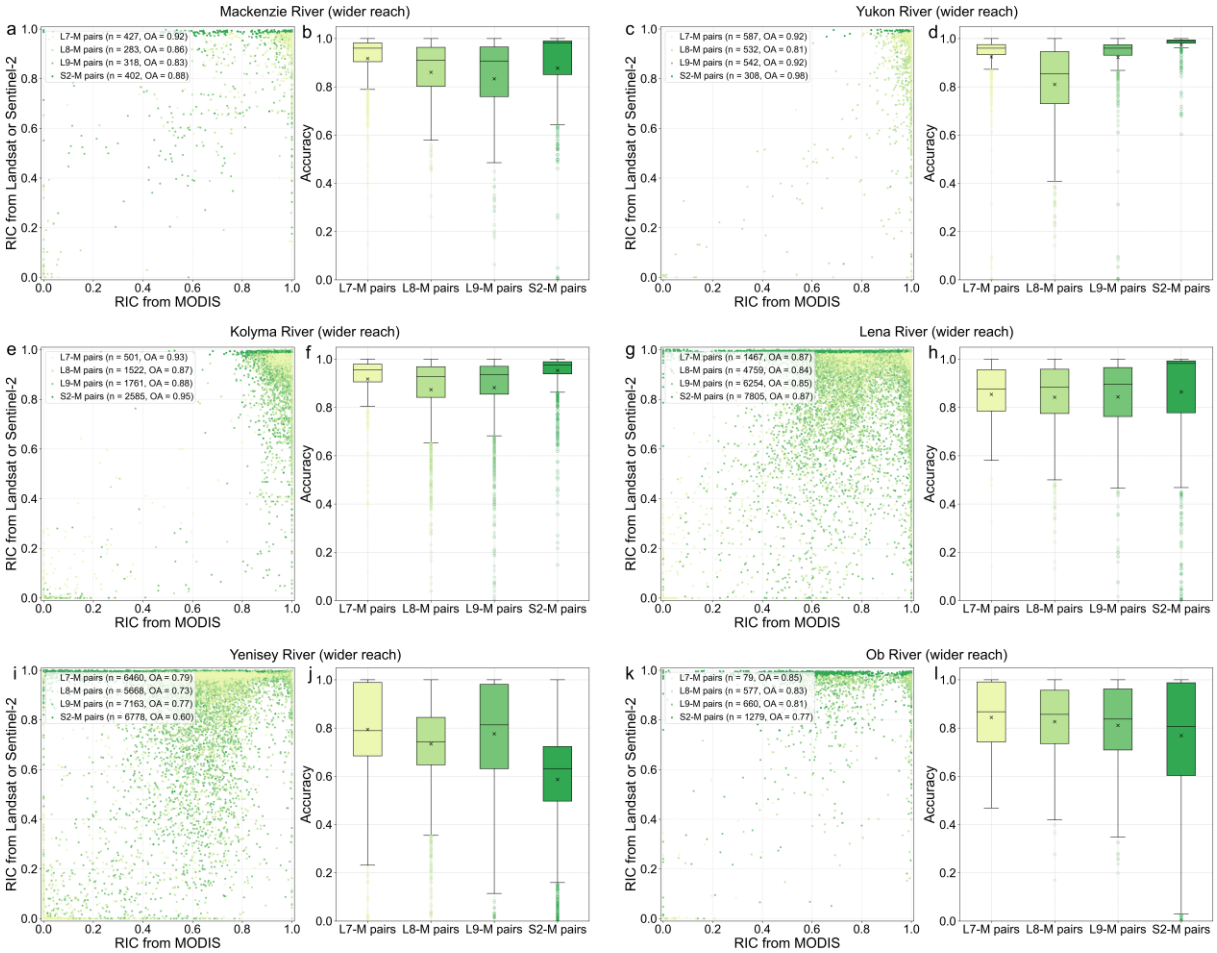

Figure II: Accuracy assessment of MODIS-derived RIC against high-resolution reference data from Landsat and Sentinel-2 across the wider reaches of six major Arctic rivers. Left panels show pixel-wise scatterplots comparing MODIS-based RIC with higher-resolution estimates from Landsat 7, Landsat 8, Landsat 9, and Sentinel-2. Right panels present the corresponding validation accuracies for each sensor pairing (L7-M, L8-M, L9-M, and S2-M). In the boxplots, the central line indicates the median, and the cross symbol denotes the mean value. Subplots a, b correspond to the Mackenzie River, c, d to the Yukon River, e, f to the Kolyma River, g, h to the Lena River, i, j to the Yenisey River, and k, l to the Ob River.

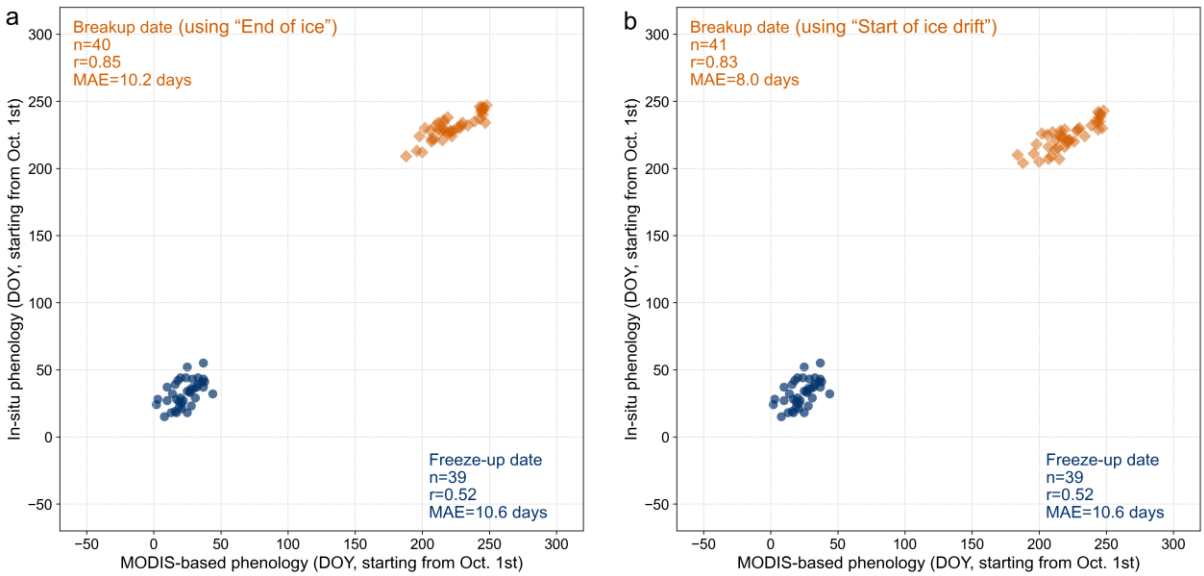

Figure III: Phenology (freeze-up and breakup dates) validation against in situ records in narrower reaches. Breakup dates were defined using "End of ice" records (a) and "Start of ice drift" records (b).

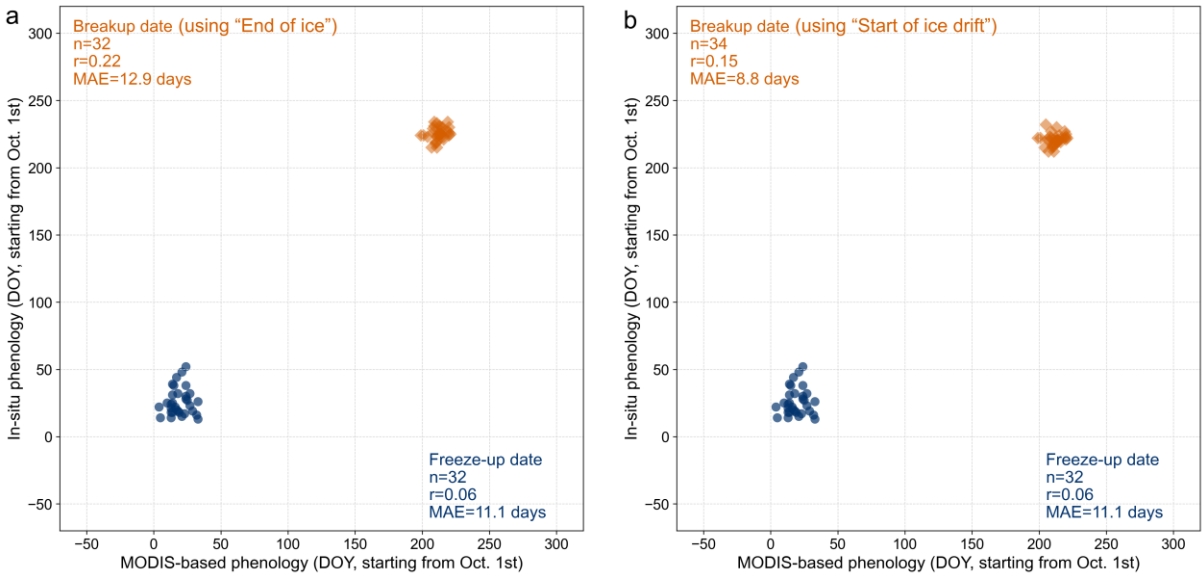

Figure IV: Phenology (freeze-up and breakup dates) validation against in situ records in wider reaches. Breakup dates were defined using "End of ice" records (a) and "Start of ice drift" records (b).

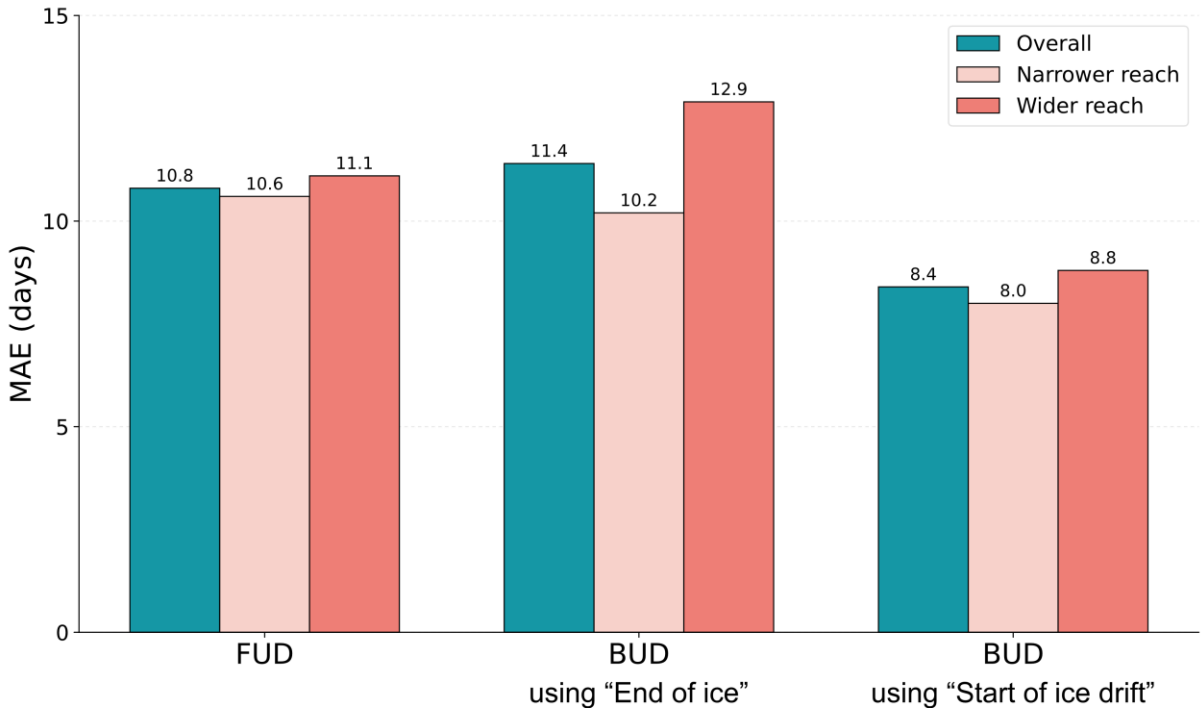

Figure V: River ice phenology MAE comparison in narrower, wider, and the whole reaches.

- As discussed above, depending on river width, mixed pixels (i.e. a pixel that observes both water and land areas) may be more common or less common. Thus, depending on how sensitive NDSI is to vegetation on the landscape, for example, it is possible that use of a single NDSI threshold to differentiate between water/land and snow/ice may be challenging or less accurate. I think additional validation of narrow reaches, as previously mentioned, could help determine if a more complicated thresholding approach is needed. Alternatively, if there is research showing that NDSI is not particularly sensitive to differences in changes in vegetation (brown vs green), that could also help support the current approach.

Response:

We thank the reviewer for this insightful comment. We agree that mixed pixels can potentially affect the effective NDSI response, especially in narrower reaches where land contributions within a 500 m MODIS pixel are more likely. In response to this concern, we have conducted the additional validation for narrower reaches (limited, overall 12%) suggested by the reviewer. The results show that the point-number-weighted mean overall accuracy of MODIS-derived

RIC is 0.81 for narrower reaches and 0.80 for wider reaches (as already discussed above, see Figs. I and II), both of which are essentially comparable to the accuracy for the full set of reaches (0.80). We also re-evaluated the river ice phenology performance separately for narrower and wider reaches. For freeze-up date (FUD), the validation results indicate broadly comparable error levels across narrower, wider, and all reaches, with mean absolute errors (MAEs) of approximately 11 days. For breakup date (BUD), the MAE based on the "start of ice drift" reference ranges from 8.0 to 8.8 days, again showing no clear degradation in the narrower reaches. Using the alternative breakup reference "End of ice", the MAE is 10.2 days for narrower reaches, which is still acceptable and even slightly lower than that for the full set of reaches (11.4 days), while the corresponding value for wider reaches is 12.9 days. Overall, these additional analyses indicate that, at the network scale of this dataset, the use of our combined thresholds ($NDSI > 0.40$, $\rho_{NIR} > 0.10$, and $\rho_{green} > 0.11$) do not lead to a clear deterioration in either RIC or phenology performance in narrower reaches.

Regarding the possible influence of vegetation conditions, we did not identify studies that explicitly isolate brown-versus-green vegetation state as an independent control on NDSI threshold performance in mixed riverine pixels. However, the existing snow-mapping literature consistently indicates that the more important controls on NDSI-based snow/ice detection are vegetation density, canopy obscuration, and mixed-pixel composition, rather than vegetation color state alone. Specifically, previous studies have shown that forest canopy is a major limitation for optical snow mapping and that algorithm performance degrades primarily because snow is partially obscured or spectrally mixed with vegetation surfaces (Hall et al., 1998; Klein et al., 1998). Likewise, the MODIS snow product algorithm applies a fixed NDSI-based framework, while additional reflectance-based screens are mainly introduced to improve performance over darker and more densely vegetated surfaces rather than to account for different vegetation color states (Hall et al., 2002; Hall and Riggs, 2007). Other studies have further shown that NDSI-based retrieval uncertainty is strongly affected by canopy structure and sensor view angle (Xin et al., 2012). Therefore, together with our additional narrower-reach validation, we believe that the current combined thresholds approach remains appropriate for this large-scale dataset, while its limitations in mixed-pixel environments are now more clearly acknowledged in the revised manuscript.

References:

Hall, D.K. and Riggs, G.A.: Accuracy assessment of the MODIS snow products. Hydrol. Process., 21(12), 1534-1547, 2007.

Hall, D.K., Foster, J.L., Verbyla, D.L., Klein, A.G. and Benson, C.S.: Assessment of snow-cover mapping accuracy in a variety of vegetation-cover densities in central Alaska. Remote Sens. Environ., 66(2), 129-137, 1998.

Hall, D.K., Riggs, G.A., Salomonson, V.V., DiGirolamo, N.E. and Bayr, K.J.: MODIS snow-cover products. Remote Sens. Environ., 83(1-2), 181-194, 2002.

Klein, A.G., Hall, D.K. and Riggs, G.A.: Improving snow cover mapping in forests through the use of a canopy reflectance model. Hydrol. Process., 12(10-11), 1723-1744, 1998.

Xin, Q., Woodcock, C.E., Liu, J., Tan, B., Melloh, R.A. and Davis, R.E.: View angle effects on MODIS snow mapping in forests. Remote Sens. Environ., 118, 50-59, 2012.

- The validation approach (Landsat 7/8/9 and Sentinel-2) relies on an NDSI threshold of 0.4 and a blue-band threshold of 0.075. All these satellites have slightly different band centers and can have slight variations instrument to instrument – thus single thresholds may not be appropriate. While NDSI is an index and is thus likely less sensitive to differences between sensors, the blue band could be more sensitive in this regard. The Harmonized Landsat and Sendinel-2 (HLS - link) dataset already somewhat solves the multi-sensor problem and could be a good dataset to switch to. Plus, HLS has data every three days at the equator and is more frequent at higher latitudes. Alternatively, their methods for dataset harmonization could be used to adjust the datasets the authors are already using. In the section beginning on line 252, they mention Landsat 7 has the highest consistency with MODIS. I am curious if the thresholds mentioned above were originally developed for Landsat 7, or one of the other Landsat sensors.

Response:

We thank the reviewer for this thoughtful and constructive comment. We agree that inter-sensor differences in spectral response, particularly for the blue band, merit careful consideration when fixed thresholds are applied across Landsat and Sentinel-2 imagery. In response, we conducted

an additional validation using the Harmonized Landsat and Sentinel-2 (HLS) product, which is specifically designed to reduce cross-sensor radiometric inconsistencies. The additional HLS-based validation (for hydrological year 2018) shows good agreement with the MODIS-derived RIC across all six rivers, with overall accuracies of 0.78, 0.81, 0.89, 0.84, 0.82, and 0.81 for the Mackenzie, Yukon, Kolyma, Lena, Yenisey, and Ob rivers, respectively (Fig. VI). In total, this comparison includes 100,842 matched HLS–MODIS grid-cell pairs, and the weighted mean overall accuracy is approximately 0.83. These results are very close to those obtained from our original validation framework (0.82) and therefore provide additional support that the adopted threshold combination is sufficiently robust, despite modest inter-sensor differences.

At the same time, we respectfully note that the study period spans more than two decades (2000–2024), and the purpose of the original validation design was to assess MODIS-derived RIC against the best available higher-resolution observations for different hydrological years. As described in the manuscript, Landsat 7 was used for the 2006 hydrological year, whereas Landsat 8, Landsat 9, and Sentinel-2 were used for the 2021, 2024, and 2018 hydrological years, respectively. This sensor-specific strategy was adopted because the operational periods and data availability of these missions differ, and it allowed us to include historical validation cases that would not be covered by a harmonized recent-era product (since 2013) alone.

Regarding the threshold choice, we clarify that the reference classification thresholds were not developed specifically for Landsat 7 alone. Rather, the approach was adopted from a previously established Landsat-based ice-mapping framework (Sojka et al., 2023) and then slightly adapted in this study. As stated in the manuscript, river ice was delineated using a fixed NDSI threshold of 0.4 together with a blue-band reflectance threshold of 0.075, where the latter was selected as the midpoint of the recommended range (0.033–0.120) to balance omission and commission errors. Thus, although the thresholds originated from Landsat-family applications, they were not tuned exclusively to Landsat 7.

Finally, we do not interpret the somewhat higher consistency of Landsat 7 with MODIS as evidence that the thresholds were inherently better suited to Landsat 7. A more likely explanation is that the Landsat 7 comparison involved fewer matched samples and was therefore less influenced by more challenging cases, such as narrow reaches and stronger mixed-pixel effects, which tend to reduce agreement between higher-resolution optical data and

MODIS. We have therefore revised the manuscript to explicitly acknowledge this issue. The revised manuscript now reads as follows:

Across the satellite platforms used for validation, Landsat 7 showed the highest consistency with MODIS, with a mean matching accuracy of 0.87 and generally narrow interquartile ranges. This was followed by Landsat 8 and Landsat 9, both of which yielded mean accuracies of 0.83. Sentinel-2, despite offering the finest spatial resolution, produced a slightly lower mean accuracy of 0.79, with greater variability in several basins compared with the Landsat-based comparisons. The comparatively higher agreement for Landsat 7 likely reflects, at least in part, its smaller number of matched samples relative to the other sensor pairings, and thus a lower influence from more challenging cases. By contrast, the lower accuracy of Sentinel-2 likely arises from its enhanced ability to detect ice within narrow river segments, where its higher spatial resolution captures a greater number of ice-covered pixels. MODIS, with its coarser 500 m resolution, is more susceptible to mixed-pixel effects in such settings, which can lead to systematic differences in ice extent and consequently biased RIC values.

Reference:

Sojka, M., Ptak, M. and Zhu, S.: Use of Landsat Satellite images in the Assessment of the variability in Ice Cover on Polish Lakes, Remote Sens., 15(12), 3030, 2023.

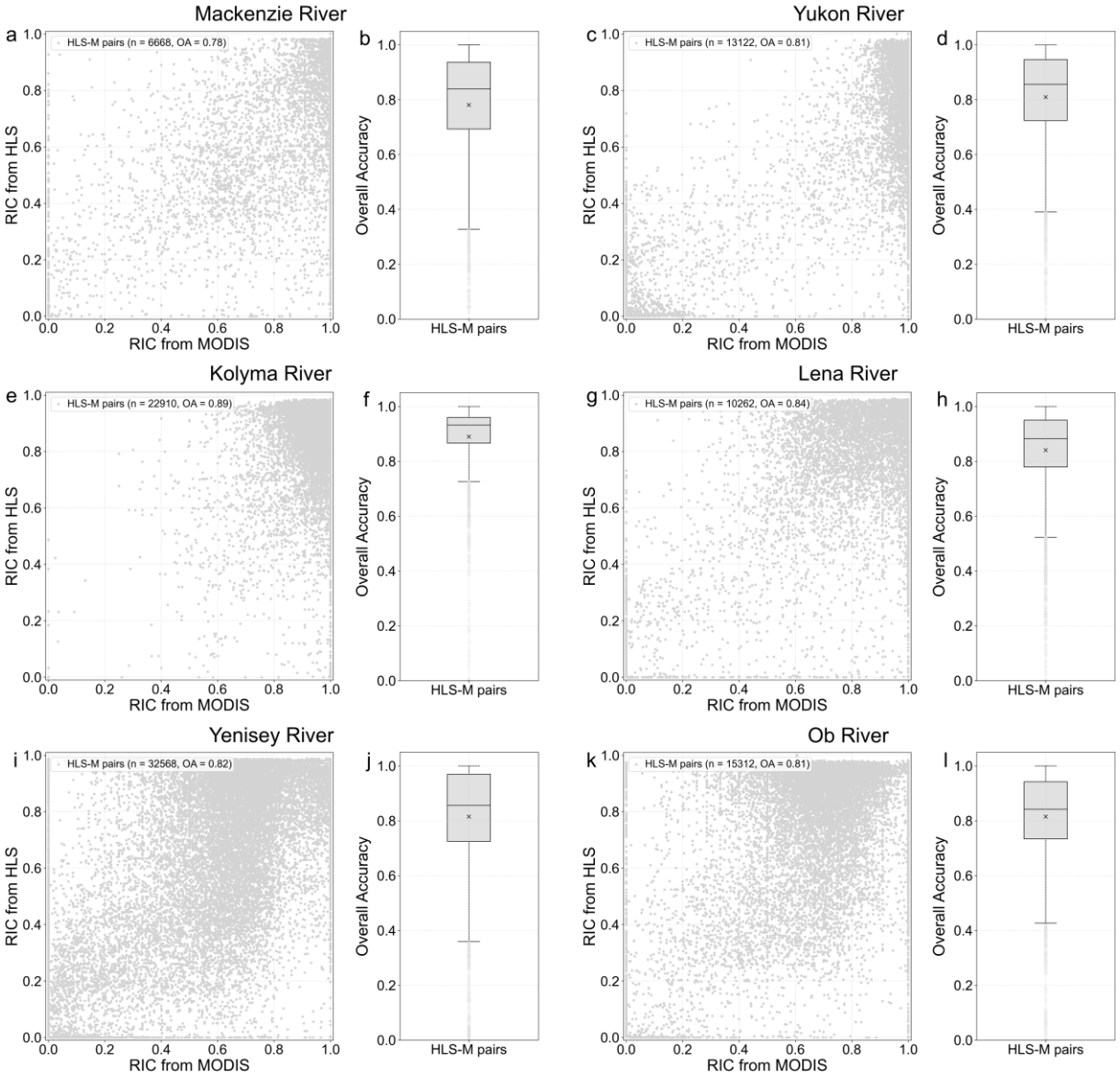

Figure VI: Accuracy assessment of MODIS-derived RIC (river ice concentration) using high-resolution Harmonized Landsat and Sendinel-2 (HLS) data across six major Arctic rivers. Left panels show pixel-wise scatterplots comparing MODIS-based RIC with higher-resolution estimates from HLS. Right panels present the corresponding validation accuracies. In the boxplots, the central line indicates the median, and the cross symbol denotes the mean value. Subplots (a), (b) correspond to the Mackenzie River, (c), (d) to the Yukon River, (e), (f) to the Kolyma River, (g), (h) to the Lena River, (i), (j) to the Yenisey River, and (k), (l) to the Ob River.

Minor comments

Overall:

- Throughout the manuscript average values are typically reported without an accompanying quantification of data spread around that mean (e.g., standard deviation or interquartile range). Adding a measure of spread would help contextualize variability in the data.

Response:

Thank you for this constructive comment. We agree that reporting only average values may not fully convey the variability of the data. In the revised manuscript, we have therefore clarified the spread of the validation results by explicitly referring to the interquartile range (IQR) shown in the boxplots in Fig. 3, in addition to the reported mean accuracies. We have also revised the related text to emphasize not only the average agreement between MODIS-derived RIC and the higher-resolution reference datasets, but also the variability across basins and sensor pairings. This revision provides a more complete description of the robustness and uncertainty of the validation results.

The corresponding revised text now reads as follows:

Figure 3 summarizes the accuracy evaluation for six major Arctic rivers, comparing MODIS-derived RIC with higher-spatial-resolution reference estimates computed using the same referenced river mask and an identical 3-km grid partition. Overall, the MODIS-based RIC achieved a mean accuracy of 0.83, with a median of 0.87 and an interquartile range (IQR) of 0.73–0.94 across all basins. Among the six rivers, the Kolyma River exhibited the strongest agreement (overall accuracy = 0.90; median = 0.93) with a relatively compact distribution (IQR = 0.86–0.97), whereas the Yenisey River—whose upstream network includes dense, lower-latitude tributaries—showed the lowest agreement (overall accuracy = 0.77; median = 0.85) and comparatively larger variability (IQR = 0.67–0.95).

Across the satellite platforms used for validation, Landsat 7 showed the highest consistency with MODIS, with a mean matching accuracy of 0.87 and generally narrow interquartile ranges. This was followed by Landsat 8 and Landsat 9, both of which yielded mean accuracies of 0.83. Sentinel-2, despite offering the finest spatial resolution, produced a slightly lower mean

accuracy of 0.79, with greater variability in several basins compared with the Landsat-based comparisons. This discrepancy likely arises from Sentinel-2's enhanced ability to detect ice within narrow river segments, where its higher resolution captures a greater number of ice-covered pixels. MODIS, with its coarser 500 m resolution, is more prone to mixed-pixel effects in such settings, leading to systematic errors of ice extent and consequently biased RIC values.

- Overall, since many of the figures have acronyms, it could be useful to make sure each acronym is spelled out in each figure caption. It is not required, but I think doing this can make the figures a little more standalone and thus easier for readers to digest.

Response:

Thank you for this helpful suggestion. We agree that spelling out acronyms in the figure captions improves readability and helps make the figures more self-contained. In the revised manuscript, we have checked the figure captions and expanded the acronyms at their first appearance in each caption where appropriate. This revision should make the figures easier to interpret independently of the main text and improve accessibility for readers.

Figures:

- Figure 1. Clear figure overall! The colors used for the Yukon and Kolyma rivers are a little challenging to see on the map.

Response:

Thank you for this helpful comment. We agree that the original colors used for the Yukon and Kolyma rivers were relatively light and could be difficult to distinguish clearly against the map background. In the revised Figure 1, we have adjusted these colors to improve contrast and visibility, while preserving the overall visual consistency of the figure. We hope this revision makes the river networks easier to identify and the figure more accessible to readers.

The updated figure is shown below:

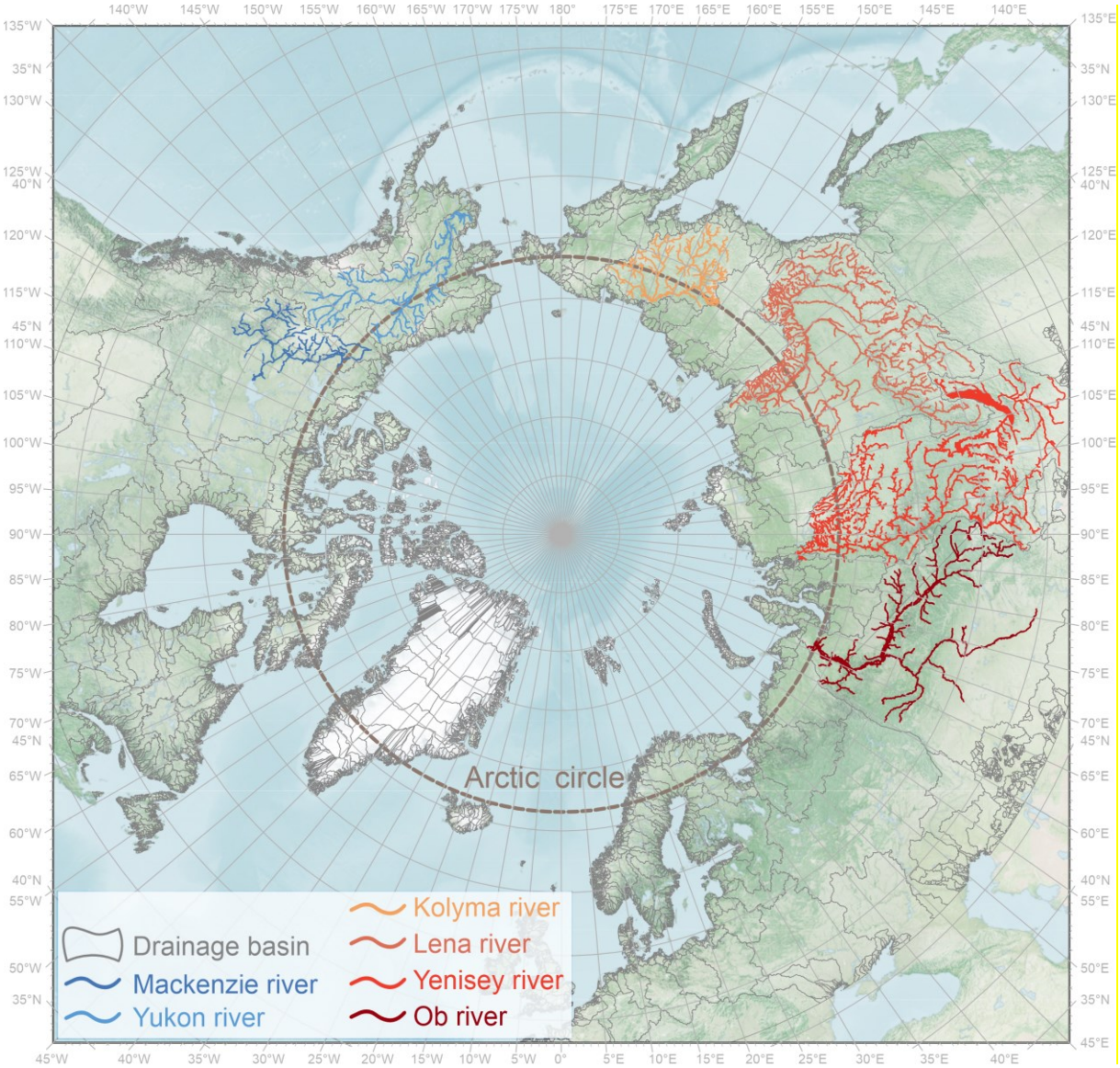

Figure 1: Study area map showing the six major pan-Arctic rivers (Yukon, Mackenzie, Ob, Yenisey, Lena, and Kolyma rivers) under the WGS 84 / Arctic Polar Stereographic projection (EPSG:3995). The background map is from NASA Earth Observatory.

- Figure 2. This flow chart was really helpful!

Response:

Thank you for your positive comment. We are pleased to hear that Figure 2 was helpful and that the flow chart clearly conveyed the workflow of the study. We appreciate this encouraging feedback.

- Figure 3. The boxplots are very clear, though it could be helpful to remind the reader what accuracy metric is being used for the y axis title. (i.e., 'Overall Accuracy', rather than just 'Accuracy'). In the scatterplots, given the color scheme and the number of points, it is a little challenging to tease out actual trends vs. differences in dot density. Also, just looking at the scatterplots, Sentinel-2 almost always has a high RIC, regardless of MODIS's RIC, yet the boxplots look pretty good for Sentinel-2. I imagine this is related to how the scatterplots were generated, but it is something to think about. The legends are also a little small to read.

Response:

Thank you for these careful and constructive comments. We agree that the y-axis label in the boxplots can be made more explicit, and in the revised figure we have changed it from 'Accuracy' to 'Overall Accuracy' to better indicate the metric being reported. We have also enlarged the legend font size to improve readability as far as possible, while keeping it sufficiently compact to avoid excessive obstruction of the scatterplot content.

Regarding the scatterplots, we appreciate this meticulous observation. The apparent pattern for the Sentinel-2 comparisons is primarily related to the plotting strategy rather than to an inconsistency in the validation results. Specifically, the scatterplots were generated using a fixed layer order for the four categories of matched points (L7-M, L8-M, L9-M, and S2-M). Because the S2-M points are shown with the darkest color, they were plotted in the bottom layer to avoid obscuring the other point pairs. As a result, the visible point distribution can give the impression that Sentinel-2 RIC values are concentrated at high levels regardless of MODIS-derived RIC, whereas the boxplots summarize the actual distribution of the matching accuracies and therefore provide a more representative quantitative comparison. To avoid this possible misunderstanding, we have clarified this point in the revised figure caption.

The updated Figure 3 is shown below:

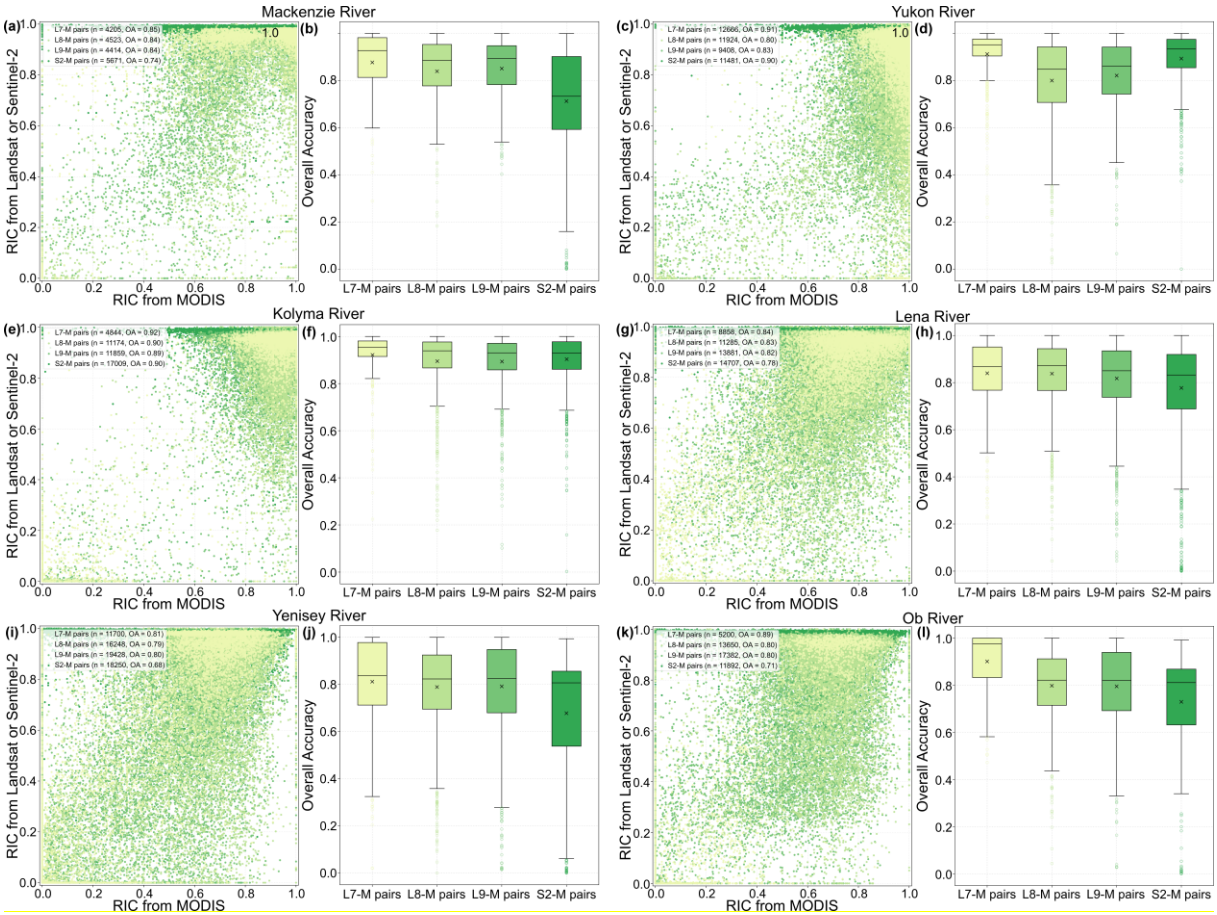

Figure 3: Accuracy assessment of MODIS-derived RIC (river ice concentration) using high-resolution reference data from Landsat and Sentinel-2 across six major Arctic rivers. Left panels show pixel-wise scatterplots comparing MODIS-based RIC with higher-resolution estimates from Landsat 7, Landsat 8, Landsat 9, and Sentinel-2. Note the four matched point categories were plotted in a fixed layer order to improve visibility, with S2-M shown in the bottom layer. Right panels present the corresponding validation accuracies for each sensor pairing (L7-M, L8-M, L9-M, and S2-M). In the boxplots, the central line indicates the median, and the cross symbol denotes the mean value. Subplots (a), (b) correspond to the Mackenzie River, (c), (d) to the Yukon River, (e), (f) to the Kolyma River, (g), (h) to the Lena River, (i), (j) to the Yenisey River, and (k), (l) to the Ob River.

- Figure 4. Nice color scheme and clear labels. Definitely not required, but if the authors wished to show patterns in breakup and freeze-up a little more clearly, they could divide each panel into two sub-panels. That way the y axis could zoom in to the correct time of year for each sub-panel (breakup vs. freeze-up).

Response:

Thank you for your comment. We agree that separating freeze-up and breakup into two sub-panels could further highlight the patterns within each seasonal phase. However, we would like

to retain the current figure layout because it presents both phenological events within the same hydrological-year framework, which more clearly illustrates their temporal separation and overall relationship. This is particularly useful in subplot (b), where the large number of data points benefits from being viewed together in a unified annual context. We believe that the current design, together with the distinct color scheme and labels, provides a clear and compact representation of both freeze-up and breakup phenology.

- Figure 5. The time series panels are very clear. To me, the scatterplot relationships look less linear, and more like a logistic regression (see Figure 5 in Yang et al. 2020). Would this benefit from assessing whether a different statistical relationship could be a better fit?

Response:

Thank you for this insightful suggestion. We agree that the RIC–SAT scatterplots in Fig. 5 are not perfectly linear, which is expected because river ice concentration is a bounded variable (0–1) and therefore tends to show saturation near the warm and cold ends of the temperature range. Following the reviewer's recommendation, we additionally evaluated a logistic fit alongside the original linear fit (Figs. VII and 5). The logistic fit better captures the curvature at the low- and high-temperature extremes, whereas the linear fit still provides a useful first-order summary of the basin-scale sensitivity of RIC to SAT over the main transition range. Importantly, this additional analysis does not change the overall interpretation: across all six basins, RIC remains strongly and negatively associated with SAT, confirming that temperature is the dominant climatic control on basin-scale river ice variability. We have therefore revised Fig. 5 and updated the manuscript text to clarify that the relationship is strongly monotonic, with nonlinear saturation at the extremes, rather than strictly linear.

The corresponding revised text now reads as follows:

To characterize the spatiotemporal behaviour of RIC, we analysed daily time series over 24 hydrological years (2001–2024) for six major Arctic rivers (Fig. 5), together with spatial distributions of mean winter RIC (Fig. 6) and basin-scale interannual anomalies (Figs. S3–S8). The time series exhibit a pronounced seasonal cycle tightly out of phase with SAT: as SAT

declines below 0 °C, basin-mean RIC rises rapidly to a stable winter plateau, followed by an equally rapid spring decline. This anti-phased relationship is strong across all basins (mean Pearson r = −0.91). Although the RIC–SAT relationship shows some sigmoidal curvature, reflecting the bounded nature of RIC and saturation near fully ice-covered and open-water conditions, the central transition range is still well approximated by a linear fit, with slopes indicating a decrease of approximately 2–3 percentage points in RIC per 1 °C increase in SAT. An additional logistic fit confirms the same overall interpretation while better capturing the curvature at the warm and cold extremes. By contrast, correlations with net solar radiation and total precipitation show consistent monotonic but weaker negative associations (basin-mean Pearson r = −0.76 and −0.43, and Spearman ρ = −0.72 and −0.45, respectively; Figs. S9–S10). High-latitude basins such as the Yukon and Kolyma consistently attain near-complete winter coverage, whereas lower-latitude tributaries show longer shoulder seasons and more frequent mid-winter interruptions.

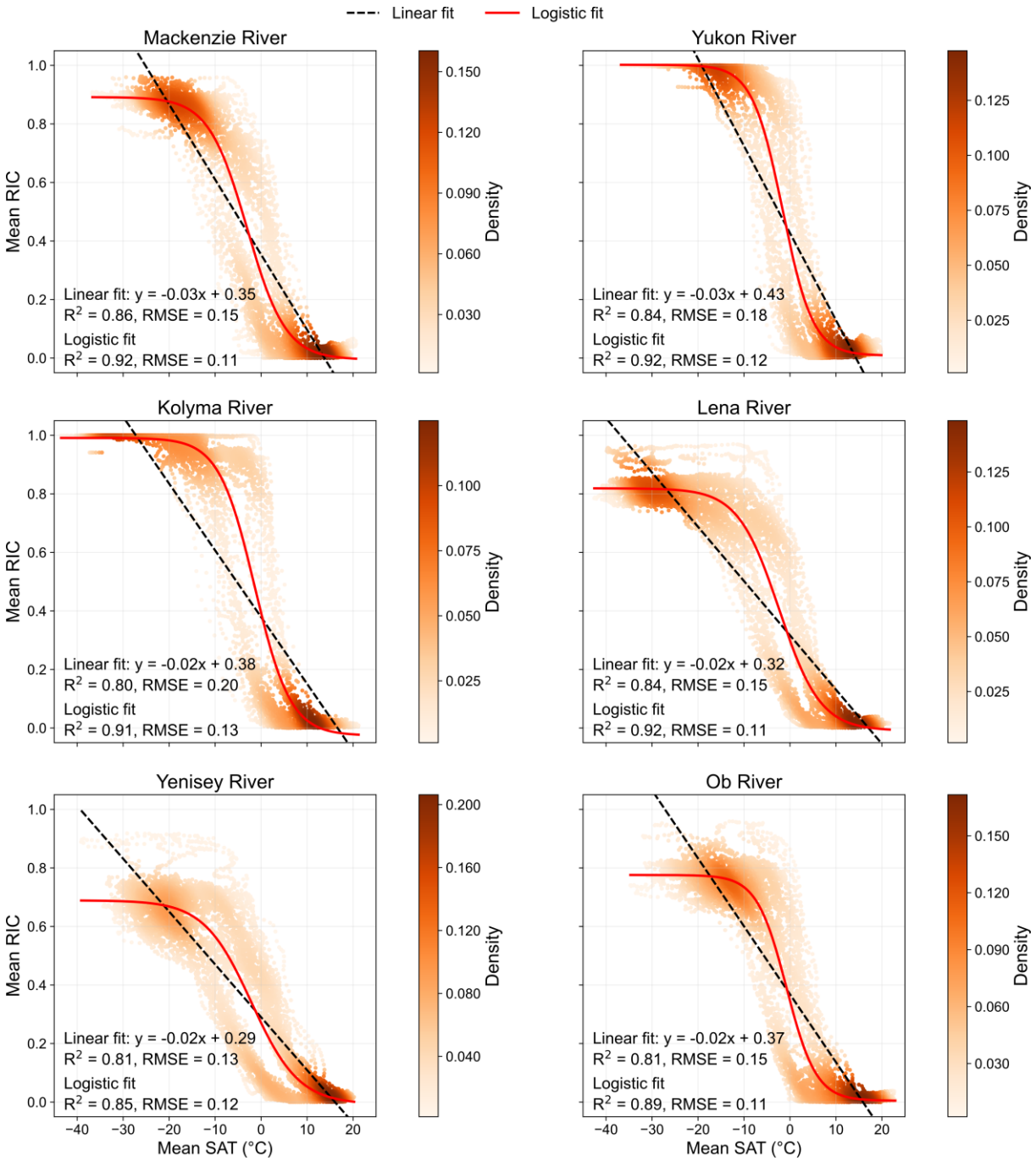

Figure VII: Phenology validation against in situ records using the uniform thresholds applied by Zhang et al (2024).

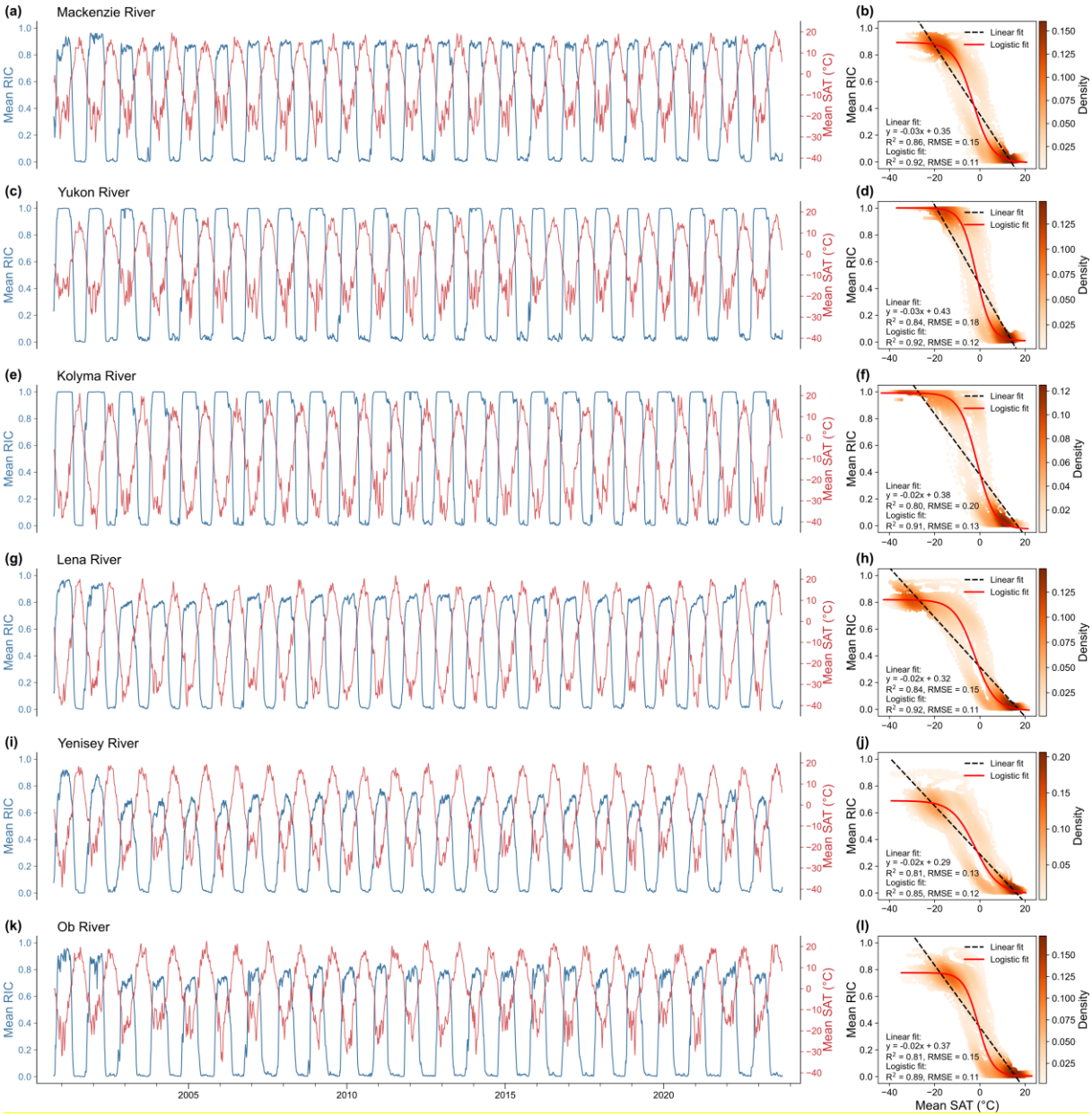

Figure 5: Long-term variations and statistical correlations between MODIS-derived RIC (river ice concentration) and SAT (surface air temperature) across six major Arctic rivers. Panels show (left) time series of mean daily RIC and ERA5-Land SAT, and (right) their corresponding correlations for each river. A 10-day moving average was applied to the SAT time series to match the temporal smoothing applied during the generation of gridded RIC products. Subplots (a), (b) correspond to the Mackenzie River, (c), (d) to the Yukon River, (e), (f) to the Kolyma River, (g), (h) to the Lena River, (i), (j) to the Yenisey River, and (k), (l) to the Ob River.

- Figures 6 – 9. I reviewed a printed color copy of the paper, and it was difficult to differentiate between continents and oceans. Perhaps this looks ok on the computer, but I do wonder if the contrast between continents and oceans could be increased on these

figures.

Response:

We thank the reviewer for your comment. We understand the concern regarding the visual contrast between continents and oceans in the printed version. In these figures, we intentionally used a white background for the oceans and a light grey background for the continents to achieve a balance between geographic context and the clear presentation of the main data layers. A stronger land–ocean contrast would increase the visual prominence of the basemap and could interfere with the readability of the scientifically relevant information, including the pixel-level colour variations, scatter-density patterns, as well as the histograms. We also selected a colour scheme that is accessible to readers with colour-vision deficiencies, and retaining a subdued basemap helps preserve the clarity of these thematic elements. For these reasons, and because the original figures can be interpreted more clearly in electronic format, we prefer to retain the current design without modification.

- Figure 10. Same comment as Figure 4.

Response:

We thank the reviewer for this helpful suggestion again. We would like to modify in this case to highlight the distinct patterns of river ice phenology across different rivers over time and improve the readability of the temporal patterns by allowing each event to be displayed over its relevant seasonal range. Accordingly, we have revised Figure 10 by splitting the original phenology panel into two sub-panels, with one showing freeze-up dates and the other showing breakup dates, while retaining the ice-duration panel separately, as shown below:

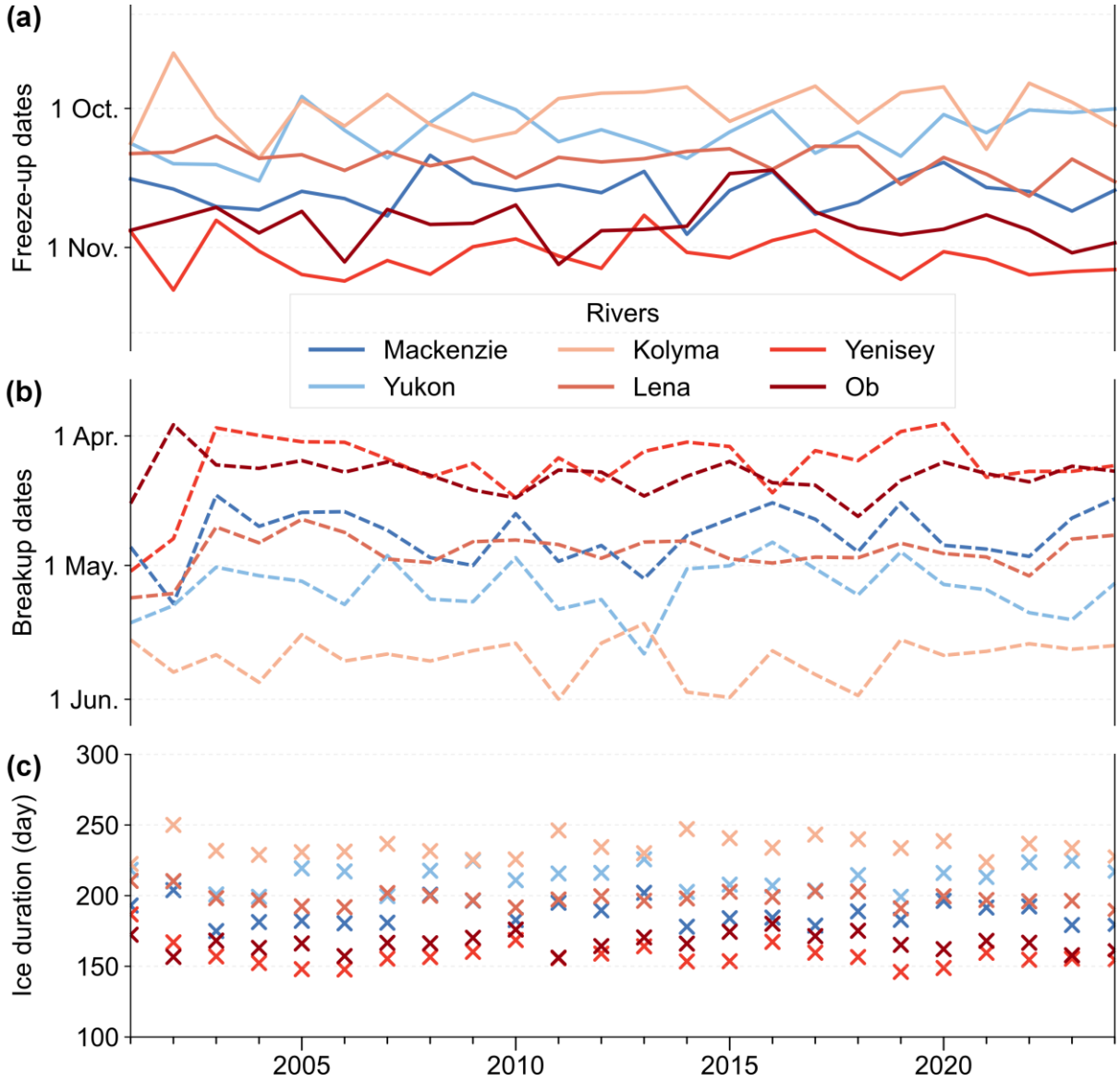

Figure 10: Annual records of river ice phenology—freeze-up, breakup, and ice duration—reported as basin-mean values across six major Arctic rivers.

- Figure 11. The ice duration plots are very clear. On the line plots, it could be helpful to add a measure of spread in freeze-up/breakup date across all 24 years, around the mean freeze-up and breakup dates. Like in Figure 1, the colors for the Yukon and Kolyma rivers may be challenging to read.

Response:

We thank the reviewer for this helpful suggestion. We agree that adding a measure of spread around the mean freeze-up and breakup curves would improve the interpretability of Figure 11. In the revised figure, we have added shaded bands representing the interquartile range (IQR)

across the 24 hydrological years around the mean freeze-up and breakup dates along each river. We have also updated the river color scheme to improve visual distinction, particularly for the Yukon and Kolyma rivers, which were less clearly separated in the previous version. The updated Figure 11 is shown below:

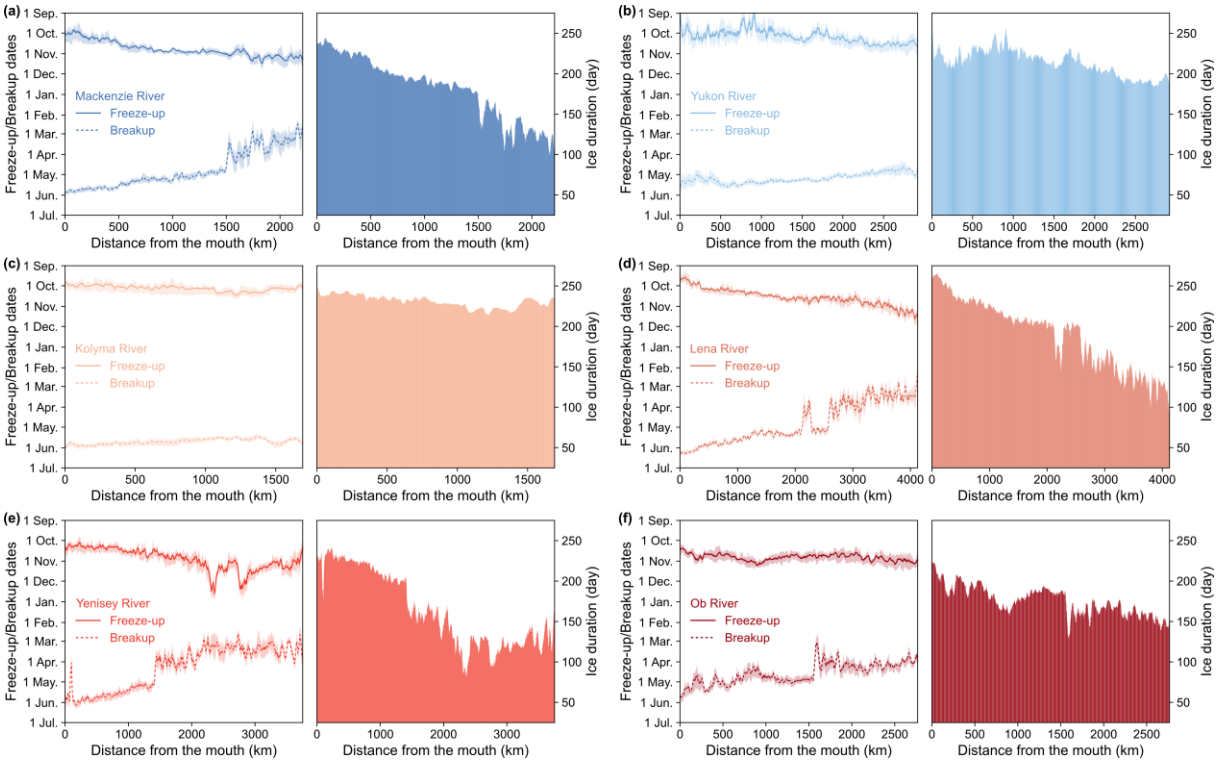

Figure 11: Spatial variations in river ice phenology along six major Arctic rivers. River ice freeze-up, breakup, and duration characteristics are shown for the (a) Mackenzie, (b) Yukon, (c) Kolyma, (d) Lena, (e) Yenisey, and (f) Ob rivers. For each river, the left panel shows the 24-year mean freeze-up and breakup dates as a function of distance from the river mouth along the main channel, with shaded bands indicating the interquartile range across the 24 hydrological years. The right panel shows the corresponding 24-year mean ice duration along the same transect. Phenology curves were smoothed using a 10-point moving average.

Introduction

- Line 72. The text mentions that existing datasets are currently deficient across the pan-Arctic region, particularly lacking those that have 'high temporal and adequate spatial resolution.' The introduction could be strengthened with a bit more discussion of why existing datasets do not meet the existing need. I.e., what sorts of questions can you answer with this new MODIS dataset that you could not answer with the coarser

temporal resolution Landsat datasets or some of the temperature-based breakup/freeze-up models?

Response:

We thank the reviewer for this helpful suggestion. We agree that the Introduction should more clearly articulate why currently available datasets do not fully meet the need for pan-Arctic river ice monitoring. Together with our community comment, we have expanded this part of the Introduction to clarify the specific advantages of an observation-based MODIS product relative to both Landsat-based datasets and temperature-driven model products.

In particular, we now emphasize that Landsat-based river ice datasets provide finer spatial detail, but their revisit interval is generally too low to consistently resolve the timing and progression of rapid freeze-up and breakup events, short-lived mid-season interruptions, and basin-wide day-to-day variability across large Arctic river systems. By contrast, the MODIS archive provides daily observations at moderate spatial resolution, which makes it possible to characterize continuous seasonal evolution, extract river ice concentration time series, and compare the synchronicity or heterogeneity of ice dynamics among major basins using a consistent observational framework. We also clarify that temperature-based phenology products are valuable for large-scale assessment, but they are not direct observations of river ice conditions and may not fully capture spatial heterogeneity or dynamic processes influenced by channel morphology, discharge, tributary inflow, and ice transport. Our MODIS-based dataset is derived entirely from daily satellite observations rather than simulated ice states, thereby avoiding uncertainties associated with model parameterization and providing an observation-based benchmark for evaluating modelled river ice phenology. These points are now stated more explicitly in the revised text. The relevant surrounding text now reads as follows:

Using hydrometric observations from Canada's gauging-station network, de Rham et al. (2020) developed the Canadian River Ice Database (CRID), providing a valuable national-scale compilation of river ice information. However, hydrologic archives remain deficient in basin-specific, long-term datasets that systematically document river ice dynamics across the pan-Arctic region. Although existing datasets provide useful information, Landsat-based products often require model-based reconstruction of daily ice conditions because of their limited revisit

frequency, while temperature-based approaches infer rather than directly observe river ice conditions. Both approaches may therefore introduce additional uncertainty, highlighting the need for a MODIS-based dataset that provides an observation-based record of river ice dynamics across major pan-Arctic river systems. To address this gap, we developed a comprehensive, multi-decadal (2000–2024) dataset based on MODIS imagery, focusing on six major pan-Arctic river systems: the Yukon, Mackenzie, Ob, Yenisey, Lena, and Kolyma rivers. Using consistent and scalable methods across entire river systems, including both main channels and tributaries, we derived detailed records of river ice concentration (i.e., the fractional areal coverage of ice within the river surface) and key phenological metrics (ice-on/off dates and ice duration). This dataset enables basin-level analysis of long-term river ice dynamics, supporting investigation into the climatic and anthropogenic factors that shape ice regimes across the circumpolar Arctic.

- Line 75. Towards the end of the introduction, river ice concentration is mentioned for the first time. However, ice concentration can mean different things to different people – spatial coverage fraction, concentration of ice in a water sample (e.g., in mg/L). It could be useful to add one short sentence describing what is meant by 'river ice concentration'.

Response:

We thank the reviewer for this helpful suggestion. We agree that the term river ice concentration may be interpreted in different ways when first introduced. In the revised manuscript, we have therefore added a brief clarification to define river ice concentration explicitly as the areal fraction of ice cover within the river surface area represented by each analysis unit. The revised texts are now read as follows:

Using consistent and scalable methods across entire river systems, including both main channels and tributaries, we derived detailed records of river ice concentration (i.e., the fractional areal coverage of ice within the river surface) and key phenological metrics (ice-on/off dates and ice duration).

Data sources

- Line 120. ‘Despite this extensive dataset, an average of 4 days per year during the 24 years lacked usable data’ – what level of cloud cover on a given day means that data from that day are not usable? I imagine that for any given pixel, more than 4 days would have missing or cloudy data, particularly in the Fall, when it is cloudier (at least in AK, where I am more familiar). A brief clarification would be helpful.

Response:

We thank the reviewer for this helpful comment. We agree that the original wording was imprecise and could be interpreted as implying that days were excluded based on a specific cloud-cover threshold. In our workflow, however, we did not discard daily MODIS images according to a predefined level of cloud cover; instead, cloudy pixels were addressed later using the cloud reclassification procedure as described in *Sect. 3.1.3 Temporal-based reclassification of cloud pixels*. Therefore, the statement that an average of 4 days per year lacked usable data was intended to refer to days with no usable MODIS optical observations in the archive, primarily due to orbital gaps, rather than to day-level exclusion caused by cloud contamination. We have revised the text accordingly to avoid this ambiguity. The updated texts are now read as follows:

Despite this extensive dataset, the MODIS optical archive still contained an average of 4 days per year without usable observations during the 24-year study period, primarily due to orbital gaps and other missing or unusable source observations (Yao et al., 2021).

Methodology

- Line 164. The state_1km band used to filter out clouds is a bit-wise band with several bits related to cloud presence/absence (cloud state, cirrus information, cloud shadow information, etc.). For reproducibility, it would be good to include which bits were used to filter out cloud-impacted pixels.

Response:

We thank the reviewer for this helpful suggestion. We agree that the specific bit information used from the MOD09GA state_1km QA band should be stated explicitly for reproducibility. In our workflow, cloud-impacted pixels were identified using bit 10 of the state_1km band, where a value of 0 indicates no cloud flag and 1 indicates cloud, following the MOD09GA internal cloud algorithm flag. We have revised the text accordingly to clarify this point:

(2) if both observations classified the pixel as cloud, as determined from bit 10 of the 1 km MOD09GA Quality Assurance (QA) band (state_1km; 0 = no cloud, 1 = cloud), it was classified as cloud (reclassified value = 10);

Validation

- Section 4.2. Validation can be a challenge when other satellite or in situ methods define breakup and freeze-up in different ways and on many different spatial scales (point, pixel, river reach) and I applaud the author's effort in compiling this validation data. I think, here, the impact (and novelty) of the paper could potentially be made clearer if it were able to quantify how much better MODIS is at detecting freeze-up and breakup relative to other methods (like the existing Landsat data or simple temperature-based models) when compared against in situ records. Given that the in situ records themselves are not perfectly comparable to the satellite data, I am open to pushback on this suggestion by the authors.

Response:

We thank the reviewer for this thoughtful and constructive suggestion. We agree that a direct comparison with other methods would be valuable for clarifying the practical contribution of the present dataset. However, we are cautious about making a strict quantitative claim that the MODIS-based dataset performs "better" than Landsat-based phenology products or temperature-based models in the current manuscript, because these approaches are not directly comparable in terms of event definition, spatial support, methodological framework, and even

the reference data used for validation.

In our study, freeze-up and breakup are derived from daily MODIS-based river ice concentration (RIC) time series using threshold-based criteria, whereas the Landsat-based dataset of Wang and Feng (2024) derives phenology from temporally sparse observations using a dual logistic regression framework and a different ice-threshold definition. Our manuscript already indicates that discrepancies relative to the Landsat-based product are likely driven primarily by differences in event definition and spatial aggregation (3 km grid cells versus river segments), rather than by the underlying ice mapping accuracy alone.

An additional source of incomparability is that the in situ reference data used in the two studies are not the same. In the present study, the in situ records were compiled from previous studies (Shiklomanov and Lammers, 2014; Shiklomanov, 2016), whereas Wang and Feng (2024) evaluated their Landsat-based phenology product using the River Ice Phenology dataset from the Global Lake and River Ice Phenology Database (Benson et al., 2000). Because the validation samples, reference definitions, and spatial collocation frameworks differ between the two studies, it is difficult to robustly quantify the degree of improvement in FUD (freeze-up date) and BUD (breakup date) detection through a direct comparison of the reported error statistics.

We also note that available in situ records are point-based and may follow locally defined criteria such as the beginning of continuous ice cover, the onset of ice drift, or the end of all ice conditions, whereas satellite products represent river segments or 3 km grid cells. In our current validation, MODIS-derived phenology shows MAEs of 10.8 days for FUD and 11.4 days for BUD against in situ records, with the BUD error decreasing to 8.4 days when compared against the onset of ice drift, further indicating that breakup validation is particularly sensitive to definition choice.

For these reasons, we believe that a strict head-to-head ranking among methods would only be methodologically robust if all products were re-evaluated against the same station-year samples using harmonized freeze-up and breakup definitions, identical spatial collocation rules, and a common in situ reference dataset. Such an exercise would be valuable, but it would require a dedicated intercomparison study beyond the scope of the present data paper. We have therefore revised the **Sect 8 Conclusion** to make this point clearer and to emphasize that the principal

contribution of the present product lies in providing a daily, observation-based benchmark for major Arctic rivers, thereby reducing reliance on temporal reconstruction from lower-frequency observations or indirect temperature-based inference.

The updated texts are now read as follows:

Taken together, these results demonstrate the value of the dataset as a consistent, basin-scale, and fully observation-based record of Arctic river ice dynamics. By relying on daily satellite observations rather than temporally reconstructed ice states or indirect meteorological inference, the dataset complements existing Landsat- and model-based products and provides an important benchmark for cross-dataset comparison and model evaluation. It also offers new opportunities to investigate the spatial heterogeneity, long-term change, and climate sensitivity of river ice regimes across the circumpolar Arctic, thereby supporting both process-oriented studies and large-scale assessments of Arctic environmental change.

Reference:

Benson, B., Magnuson, J. and Sharma, S.: Global Lake and River Ice Phenology Database, Version 1 [Data Set], National Snow and Ice Data Center, Boulder, Colorado USA, 10.7265/N5W66HP8, 2000.

Shiklomanov, A.I. and Lammers, R.B.: River ice responses to a warming Arctic—recent evidence from Russian rivers. Environ. Res. Lett., 9(3), 035008, 2014.

Shiklomanov, A.I.: Data of river ice timing and thickness for selected river gauges in Russian pan-Arctic over 1970-2012, Arctic Data Center, doi:10.18739/A22Z12Q02, 2016.

Wang, X. and Feng, L.: Patterns and Trends in Northern Hemisphere River Ice Phenology from 2000 to 2021, Remote Sens. Environ., 313, 114346, 2024.

Discussion

- Line 311. There is a space missing after ‘S10).’

Response:

We thank the reviewer for carefully noting this typographical error. The missing space after "S10)" has been corrected in the revised manuscript.

Data availability

- Links to data and code all work. Data is well described, including units.

Response:

We thank the reviewer for this positive assessment. We appreciate the reviewer's recognition that the data availability and documentation support the transparency and usability of the dataset.

Once again, we sincerely appreciate your meticulous review and insightful comments. We hope that our revisions effectively address your concerns and further improve the clarity, rigor, and overall quality of the manuscript.