Response to Reviewer 2

1. Major Comments
    1.1. We thank the reviewer for their diligence in going through the literature review and upon revaluation we found a few places where citations were misattested or confusing as to what claim they were supporting. The introduction has been revised to address these issues.
    1.2. The reviewer also mentions a lack of discussion of model physics. We added some discussion of CICEv5, used in CESM2 LENS and CESM2 S2S, including some known assumptions built into the model that may be affecting its performance. Specifically that it does not couple ocean waves to sea ice.
        1.2.1. We first establish that CESM2 tends to struggle to accurately represent sea ice processes on subseasonal and shorter time scales and then identify physical processes in the atmosphere that CESM2 is capturing but seems to not be affecting the sea ice. For example, in our case studies we see that CESM2 S2S tends to capture the surface lows and wind patterns associated with VRILEs present in ERA5 data, but not the associated sea ice loss. These winds likely affect the sea ice through direct transport of the sea ice floes, transporting warmer air poleward, driving ocean waves that can mechanically break up the sea ice, and driving vertical ocean mixing that contributes to bottom melt. Why those processes aren't affecting the sea ice in CESM2, apart from the known fact that neither CESM2 nor any other model couples sea ice to ocean waves, is beyond the scope of this paper.
        1.2.2. We better emphasized the physical processes discussed above in the introduction and reiterate then throughout the paper.
2. Specific Comments
    2.1. P1, L 14: For a manuscript tracking modern sea ice, citing a "twice the rate" statistic from 2007 and 2012 is quite outdated. Even with the 2017 citation included, this figure may be dated for a manuscript addressing modern sea ice — more recent estimates put Arctic warming well above twice the global rate. Consider updating with a more representative and current citations.
        2.1.1. Restated to say "more than twice" and included a more recent citation from England, Mark R., et al. (2021)
    2.2. Pg 1, L 18: The citation of Durkalec et al. (2015) here appears to be a misattribution. Durkalec et al. (2015) is a qualitative study focused on the health, safety, and cultural well-being of Inuit community members experiencing sea ice decline. It does not provide data or conclusions regarding a general increase in the Arctic human population due to sea ice loss. Please find a more suitable citation.
        2.2.1. Citation changed to: Heleniak, Timothy. "The future of the Arctic populations." *Polar Geography* 44.2 (2021): 136-152. With the Durkalec

et al. (2015) attestation moved to be a source for the needs of Arctic stakeholders.

2.3. Pg 2, L31: grammar error. The citation, DeRepentigny et al. (2020), should all be in parenthesis.

2.3.1. Corrected

2.4. Pg 2 ,L 36: Who is "they?"

2.4.1. Clarification added

2.5. Pg 2, L45: This statement "March sea ice variability is driven by radiative feedbacks of clouds and water vapor (Luo et al., 2017), surface albedo, surface winds, and oceanic heat transport, while September sea ice interannual variability is primarily driven by surface albedo and fluctuations in temperature advection (Olonscheck et al., 2019)" oversimplifies what drives sea ice variability and does not mention the regional differences and other factors like local wind forcings. The citations used for these types of overarching statements are not used correctly. For example, Luo et al. (2017) is cited to support a general Arctic-wide claim about March variability, but that paper is actually about winter ice decline in the Barents-Kara Sea specifically. And the claim that September variability is "primarily driven by surface albedo" doesn't match Olonscheck et al. (2019) where the citation clearly states "Arctic sea-ice variability is primarily driven by atmospheric temperature fluctuations," not surface albedo. There needs to be more comprehensive literature review for these generalizations and not relying on single studies for broad claims.

2.5.1. That paragraph has been significantly edited to address these concerns.

2.6. Pg 3, L63: grammar error: "One"...and then "are" should be "is"

2.6.1. Corrected

2.7. Pg 3, L76-82: From this statement it is unclear to me what the purpose of this study is. The authors used just one model simulation out of many available, picked because it looked "average," not because it gave good results. But testing only one run doesn't tell you if the model actually works well at predicting short-term ice changes — you need to check several runs to know if that's a real pattern or just luck from that one run.If the purpose is to evaluate the model's ability to simulate subseasonal variability, shouldn't multiple individual members be evaluated in order to get a statistically robust sample. If not, the authors should clarify why individual members' daily data, stated as archived in the manuscript, weren't used to test multiple members rather than just one.

2.7.1. Clarified that the single ensemble member was chosen to act as a proxy for the Large Ensemble mean by selecting the member with the highest correlation in monthly mean SIE to CESM2 LENS.

2.8. Pg 3, L76-77: The manuscript uses the NSIDC Bootstrap product as ground truth without discussing its known retrieval limitations. Passive microwave sea ice concentration algorithms, including Bootstrap, have documented uncertainty in the marginal ice zone (roughly 15-80% SIC, season and region dependant), driven by factors such as thin ice, slush, and liquid water fraction in the snowpack or snowmelt. Separately, retrievals are known to systematically underestimate

SIC during melt season, since melt ponds contaminate the reference values the algorithms are calibrated against, with error increasing as SIC decreases. Given that several VRILE cases in this study occur during summer melt conditions (e.g., July 2007, August 2012) and that VRILE identification depends on detecting concentration changes in exactly this intermediate range, this is a relevant source of uncertainty that should be acknowledged. The authors should discuss how retrieval uncertainty in NSIDC data may affect the quantitative comparison to CESM2, particularly the magnitude of the reported VRILE size discrepancy.

2.8.1. VRILES are identified by changes in sea ice extent, not concentration and the size discrepancies discussed in this section of the paper are also with respect to SIE. A full discussion of the data and methods used for VRILE identification can be found in Cavallo et. al. (2025).

2.8.2. We have added additional discussion of the NSIDC dataset in the data and methods section.

2.9. Pg 3, L88-92: Sea ice physics are non-linear. The Arctic climate forcing, atmospheric CO_2 concentrations, ocean heat content, and the proportion of thick multi-year ice vs. thin first-year ice (and others) were substantially different in 1960 compared to 2020. Selecting years based on matching sea ice extent does not establish that the underlying variability-generating physics are comparable between these two periods. Since this comparison underlies the paper's interannual variability statistics (PSD, interquartile range, spatial SIC variability), please explain why these statistics should be expected to transfer across two periods with such different background climate/environmental states.

2.9.1. Model year 1960 in CESM2 LENS has no relationship to the conditions on Earth in the calendar year 1960. More detail about the CESM2 LENS model configuration is included, including a discussion of anthropogenic forcings, in the data and methods section to prevent this confusion.

2.10. pg 4., L 99-103: Regridding coarse model data onto a finer satellite grid is not a good verification approach. In practice, it introduces smearing artifacts in areas like the marginal ice zone which is critical to capture when looking at subseasonal capabilities. Rather the fine observations should instead be upscaled to the model's resolution. Either the author should try to redo the setup to see how this affects the outcome or acknowledge the practical limitations of this approach.

2.10.1. While we appreciate the reviewer’s concerns, there wasn’t feasibly another way to go about the regridding given the available data and tools. CESM2 LENS’s native grid is 1x1 degrees. However, the NSIDC sea ice concentration fields are saved on a rotated north polar stereographic projection with points spaced 25 km apart. Adapting these two datasets so that they are on the same grid and can be subtracted is much more involved that simply interpolating CESM2 LENS data between points. The tool that ended up being the most effective for this task was the scipy interpolate package, “RegularGridInterpolator”. While dealing with the rotation and map projections were notable, the real difficulty came from

the fact that either dataset is irregularly spaced on the other's native grid and the RegularGridInterpolator tool works best when it is interpolating corse, irregular data onto a finer, regular grid.

2.10.2. We expanded our discussion of the smoothing effects seen in the composite differences in Figure 4 and added detail to our description of the regridding process to the data and methods section.

2.11. Pg 4., L 103-107: Two things need clarification. First, since VRILE locations are found using the same gridded data that was regridded earlier (see comment on L99-103), the blurring from that step could shift where the center-of-mass calculation places the event. The authors should quantify how much error this introduces. For example, by comparing VRILE locations derived from the regridded data against locations derived from the model's native grid, and reporting the resulting shift in center-of-mass position. Additionally, percent change in sea ice coverage is used as a proxy for mass, but no formula or example is given, and the phrase is ambiguous. For example, the definition could refer to a percentage-point change in concentration or a relative percent change, which would produce different weightings. Please specify the exact definition used.

2.11.1. VRILE locations are found on the native grids. Added this specification to the text and clarification to the center of mass calculation.

2.12. Pg 4, L 124-129: The lead times used across the four VRILE case studies range from three to eight days (Table 1), with the manuscript noting that using earlier forecasts "yielded highly unrealistic results" for the 2012 and 2016 cases. In S2S forecasting, lead time directly affects atmospheric predictive skill, so comparing cases initialized at inconsistent lead times makes it difficult to distinguish limitations of the model's physics from limitations of atmospheric predictability at the lead time used. Please either standardize lead time across cases or discuss explicitly how this variation affects the comparability of conclusions drawn from the four cases. Additionally, please clarify the reasoning behind the statement that a 2 July 2007 initialization "does not provide enough time" to evaluate the 3 July event. As written, this is unclear. Is the constraint that the method requires a five-day antecedent SIC change to identify and locate the VRILE, which a one-day lead cannot support, rather than an issue with single-day forecast validity itself?

2.12.1. The July 2 lead time note has been removed from the text and replaced with a more physically driven discussion.

2.12.2. Since the CESM2 S2S reforecasts are initiated every Monday, standardizing lead time across cases wasn't a viable option. Additional discussions of how the different lead times affect the analysis have been added for each case.

2.13. Pg 5, L131-132: same comment regarding regridding from coarse to fine (see comment for pg 4., L 99-103)

2.13.1. Please see response 2.10.1

2.14. Pg 5, L134-255: Several methodological issues affect most of the results in this section. This is presented in both the individual figures and the overall conclusions drawn throughout Section 3. Rather than repeat these points I have referenced earlier in section 2, they are summarized here:

2.14.1. Figures 2 through 5 are all derived from a single CESM2 LENS ensemble member, with no evaluation of whether results would differ using other members see comment on Pg 3, L76-82). This concern is compounded by comparing that member's 1960-2000 period to NSIDC's 1980-2020 observational record, despite differing background climate/environmental conditions between the two periods, without demonstrating the relevance of this comparison (see comment on Pg 3, L88-92). The author should provide a more relevant literature review that characterizes the retrieval uncertainty for the ground-truth data used, particularly in the marginal ice zone and seasons in which this study is presented. It would be good to acknowledge whether melt ponds or wet snow was likely present in regions that may have introduced larger retreival errors.

2.14.1.1. Please see responses 2.7.1 and 2.9.1

2.14.2. The four case studies in Section 3.1 raise a related concern, each using a different forecast lead time (3-8 days), with two cases showing that shorter lead times produced unrealistic results, previously raised as a possibility that the cases shown are not representative (see comment on Pg 4, L124-129).

2.14.2.1. Please see response 2.12.2.

2.14.3. Last, the observational uncertainty noted in the comment on Pg 3, L76-77 applies here as well, since the NSIDC data used as the benchmark throughout this section carries its own unaddressed retrieval uncertainty in exactly the concentration range and season most relevant to VRILE detection. These caveats around using passive microwave data as ground truth do not appear to have been addressed anywhere in the study design. The discrepancy between CESM2 and observations shown in this section is clearly substantial, but given all of the above, the paper's central quantitative claim that CESM2 underestimates rapid ice-loss events by roughly an order of magnitude is less certain than presented. It is the precision of the magnitude that is of interest, not the general conclusion that CESM2 underrepresents short-term ice loss, as this is already consistent with prior literature on the model's known limitations. It is recommended that the author provides the model comparison against the observational uncertainty range so it is more clear where the gaps are significant.

2.14.3.1. Please see response 2.8.1 and 2.8.2

2.15. Pg 9, L 269 and 270: The abstract already acknowledges that models underrepresent subseasonal sea ice processes, but this statement goes further, suggesting that better representation of synoptic-scale processes would fix the forecasting shortfall. First, is the term "synoptic-scale" processes meant to be

"regional or mesoscale" when referring to the detail that is missing? The use of "synoptic-scale" doesn't match the scale identified in the paper's own explanation n Section 3.2 (L251-255). It seems the authors suggest the limiting factor is grid resolution relative to much finer scales, ice floes (~100 m) and low-level jets (~17-23 km), not synoptic scale systems (~1,000+ km). While stated on pg. 8 L255 that this is a parameterization issue and consistent with "representation" being part of the fix, the scale named in section 4 doesn't match the scale the paper's own diagnosis points to. The authors should clarify which scale they mean, since synoptic-scale representation and floe/jet-scale parameterization are different factors that require different solutions to evaluate the model correctly.

2.15.1. In this context, synoptic scale refers to the size and timing of sea ice loss. This section has been reworded to eliminate the confusion.