Response to referees for
"Brief Communication: The Interdecadal Pacific Oscillation is not a Reliable Indicator of the Timing of an Ice-Free Arctic"
by Christopher Wyburn-Powell and Alexandra Jahn

RC1: 'Comment on egusphere-2026-2082', Anonymous Referee #1, 05 May 2026 reply
This short paper is an important contribution. It takes as a starting point the work of Screen and Deser, which suggested with a single model that the IPO could be a useful constraint on the timing of an ice-free Arctic. As with all such potential constraints, it is valuable to test their robustness across multiple models. This submission tests the constraint across 8 models and finds that it is not robust. I support publication after revision and/or suitable responses to the following points:

We thank the anonymous referee for their detailed comments and suggestions, which we believe we can fully address in a revised manuscript (see below).

Line 60: I don't understand the rationale for the ensemble pooling. The text states "to incorporate multiple initial conditions", however, as far as I understand it, these are not independent runs but multiple samples from the same ensemble member. Unlike a distinct ensemble member, which is free to evolve differently to all other members, these pooled samples are not independent and are simply offset by a few years. This process seems to artificially inflate the ensemble size, without providing independent samples, and needs better justification.
We agree that the ensemble pooling could have been better motivated and explained. We will rectify this in the revised manuscript. In short, we use ensemble pooling to expand the analysis to multiple initial mean sea ice states beside the specific 4.7 million km^2 initial mean state used by Screen and Deser (2019), as our analysis showed that using a different starting mean state (e.g., 4.2, 4.3, 5.2 million km^2) changes the results even for the CESM1 data used in Screen and Deser (2019). As there is no physical reason that the response should be dependent on the exact initial mean state of 4.7 million km^2, the ensemble pooling using multiple initial mean states makes the results more robust (in particular in light of the uncertainties in the exact magnitude of the satellite-derived mean sea ice extent). In the revised manuscript, we will add the following to motivate the ensemble pooling:
"This [ensemble pooling] removes the sensitivity to a specific 4.7 million km^2 mean state as for Screen & Deser (2019), by pooling the same analysis over 11 mean states from 4.2 to 5.2 million km^2."

Additionally we will change the language in the methods section to be more specific to refer to multiple initial "mean states" rather than "initial conditions". Please also see our response to the next comment where we address the ensemble size considerations with regards the pooling methodology.

Line 76: Does this increase in the percentage of LEs showing significant divergence come from the alternative IPO definition and/or the pooling of initial states? If the latter, I'm concerned that significance is inflated artificially and has limited physical meaning. Does the significance testing assume each sample is independent? They aren't (see above). Besides, regardless of the significance or not of the divergence, the more important point seems to be that the divergence is small around the time of first ice-free conditions, which is independent of the pooling (as the samples still have the same timing of ice-free conditions). Basically, I'm wondering if the pooling is necessary, even if it could be justified.
This divergence is from the pooling methodology, however, the significance test is performed on the number of ensemble members, not the repeated sampling of those ensemble members. Thus, the significance is not affected by the pooling. We have added the following sentence to the methods for clarification: "***Significance testing uses the number of LE ensemble members as independent samples, not the multiple samplings of the same ensemble member."***

Section 3.2: I think this section could be clearer. There are lots of kind-of "what if" statements and the key result – which I think is it that there is no consistent lagged relationship between IPO and ice-free conditions across models (in fact, in most models there is no relationship at all) – gets a bit lost. I think this point could be made more succinctly and clearly, alongside some simplification of Figure 3 (see comment below).
We agree that the main point in section 3.2 is a little lost. We will simplify Figure 3 as suggested (also see below), which will help to retain focus on the main results of this section. Further we have tightened up the more speculative portions of this section to not open discussion of questions not raised in the introduction or pertinent to the conclusions.

Figure 3: It's unclear what value panels I-P add, and they are not directly referred to in the text. If I understand correctly, the correlation shown in panels I-P can already been inferred from the blue lines in panels A-H. Panels A-H include a second metric for ice-free, which is not used (or introduced) before. Is this necessary? What is the added value of the red line? In such a short note, I'm in favour of parsimony: keeping things as simple as possible.
We agree that panels I-P are not well-motivated, in the revised manuscript we will remove panels I-P. Additionally, we consider there to be minimal utility to the red lines in A-H and will therefore remove them so as not to distract from the main findings.

Conclusions: The lack of robustness of the constraint across models is an important result, but there is scope for deeper consideration of the causes of this lack of robustness, its implications, and potential ways forward. The IPO constraint can only ever reduce the model uncertainty due to internal variability. Its utility is dependent on the timeframe for an ice-free Arctic, which also depends on the simulated forced response. In this regard, it might not be appropriate to consider all models as equal. There is no effort to critique if any of models have (un)realistic forced responses, for example, by looking at their historical trends or sea-ice sensitivity (sea-ice loss per degree of global warming). At the very least, this requires some thought and discussion. A more sophisticated approach might be to first constrain the model spread in the

forced response, before considering if internal variability can additionally constrain the spread around this forced response.
In using the 8 available large ensembles (LEs), we recognize that many of the LEs have biases relative to observations in terms of the forced response, internal variability, or in the case of MIROC6 a weak Pacific-Arctic teleconnection. However, none of these models are extreme outliers for CMIP5 or CMIP6. To the authors' knowledge, if we were to select some of these LEs to use only the more 'realistic' models on certain metrics, we would not see an increase in the consensus of the IPO as a good/poor indicator for the timing of an ice-free Arctic. Nonetheless, we have added the following sentences to the conclusions section that while there is no agreement on the ice-free timing based on the IPO across models, further work is needed to determine which model's response to the IPO, if any, is most realistic in capturing the IPO's teleconnections to the Arctic.
"Further work is required to identify model realism of the Arctic response to the IPO, beyond generalized realism metrics such as forced response which do not correlate with the IPO-Arctic signal. Refining model selection based on IPO-Arctic realism may lead to a greater consensus on whether or not the IPO is a reliable indicator of the timing of an ice-free Arctic."

To further directly address your valid points raised above, we have added a column to Table 1 with dSIA/dGMST forced response values as calculated in the overview paper about Arctic sea ice in CMIP6 by Notz & SIMIP Community (2020). We will also add a discussion in the introduction which motivates the inclusion of all 8 LEs selected, despite variations in realism between the models on various metrics.
"The eight large ensembles vary across common realism metrics such as forced response (Table 1), but the specific IPO-Arctic response has limited research to constrain a sample of large ensembles."

Other minor points:

Typo line 18: "showed the IPO has been shown"
Thank you, will be corrected in the revised manuscript.

Figure 1: might be helpful to list the number of members on the titles
The eight LEs are considered equally valid as they have 30 or greater ensemble members, hence we consider displaying the number of ensemble members would highlight these differences which do not impact the results as significance is tested relative to the number of ensemble members. We note that the full list of LEs with ensemble members are listed in Table 1.

Line 134: "at today's lead times" confusing and probably not needed. The sentence makes (more) sense without this bit.
We agree that the phrase was unnecessarily ambiguous. We have reworded this sentence in the revised manuscript to reference the mean state range used in the pooled methodology.

"Therefore, there is no consensus of the utility of the IPO for determining when consistently ice-free conditions will occur when assessed from September mean sea ice extents between 4.2-5.2 million km^2."

Citation: https://doi.org/10.5194/egusphere-2026-2082-RC1

References:

Notz, D. and SIMIP-Community: Arctic Sea Ice in CMIP6, Geophysical Research Letters, 47, https://doi.org/10.1029/2019GL086749, 2020.