The authors are very grateful for the reviewer's time and effort in providing comments. Their work has greatly improved our manuscript. Our responses to each comment are included in black below.

General Comments

I wonder if even a data assimilation procedure that perfectly handled the non-Gaussian aspects of the problem would still result in a biased ensemble mean, since the distributions would necessarily be skewed. A discussion on what a "perfect" data assimilation procedure (i.e. something closely approximating Bayes theorem) would produce would be very helpful for the reader.

A sentence was added in the conclusion addressing this comment (line 405).

"This would include using distributions for the prior PDF and the observation likelihood that are similar to the observation error distribution and consider the bounds more appropriately (e.g., a truncated Gaussian distribution)."

A good example supporting the reviewer's hypothesis occurs when actual ice concentration is 0. Any ensemble estimate except for all 0's will be biased relative to the truth. However, it is less clear whether a perfect DA system can be unbiased compared to the observations. Since SIC is doubly bounded, generated observations would be inherently biased near either bound of zero or one. While the authors do agree that this would be helpful for the reader, it is a bit out of the scope of this work to speculate what a "perfect" data assimilation procedure would entail. The authors believe how to handle non-Gaussian situations is still an active area of research in the data assimilation community.

Also, I am interested to see if the ensemble median SIC is less biased than the ensemble mean and may provide a "better" measure than the mean for skewed distributions (though I admit the definition of "better" is not completely clear). This would not require additional experiments to be performed, just recomputing the bias scores by using the ensemble median, instead of the mean.

To address this comment the authors included several figures displaying the difference between using the ensemble median versus the ensemble mean. Figures 1 and 2 show the differences in the total Arctic sea ice area, sea ice volume and snow volume when using the mean versus the median (this is the same as figure 3 in the manuscript). The differences are quite small and not really noticeable. The same can be said about the spatial bias plots that were computed in the manuscript (Figures 3 and 4). The bias patterns and magnitudes are similar whether you use the median or the mean for computing the differences compared to the truth. The similarities between using the mean versus median is likely linked back to the ensemble spread being pretty small for the difference experiments and suggests that many of the distributions are approximately normal so that the mean and median are nearly the same. (see Figures 5 and 6 near the bottom of this document). The small ensemble spread likely means there are no extreme outliers effecting the computation of the mean. Due to the similarities, the authors have chosen to use the mean in this manuscript (line 185).

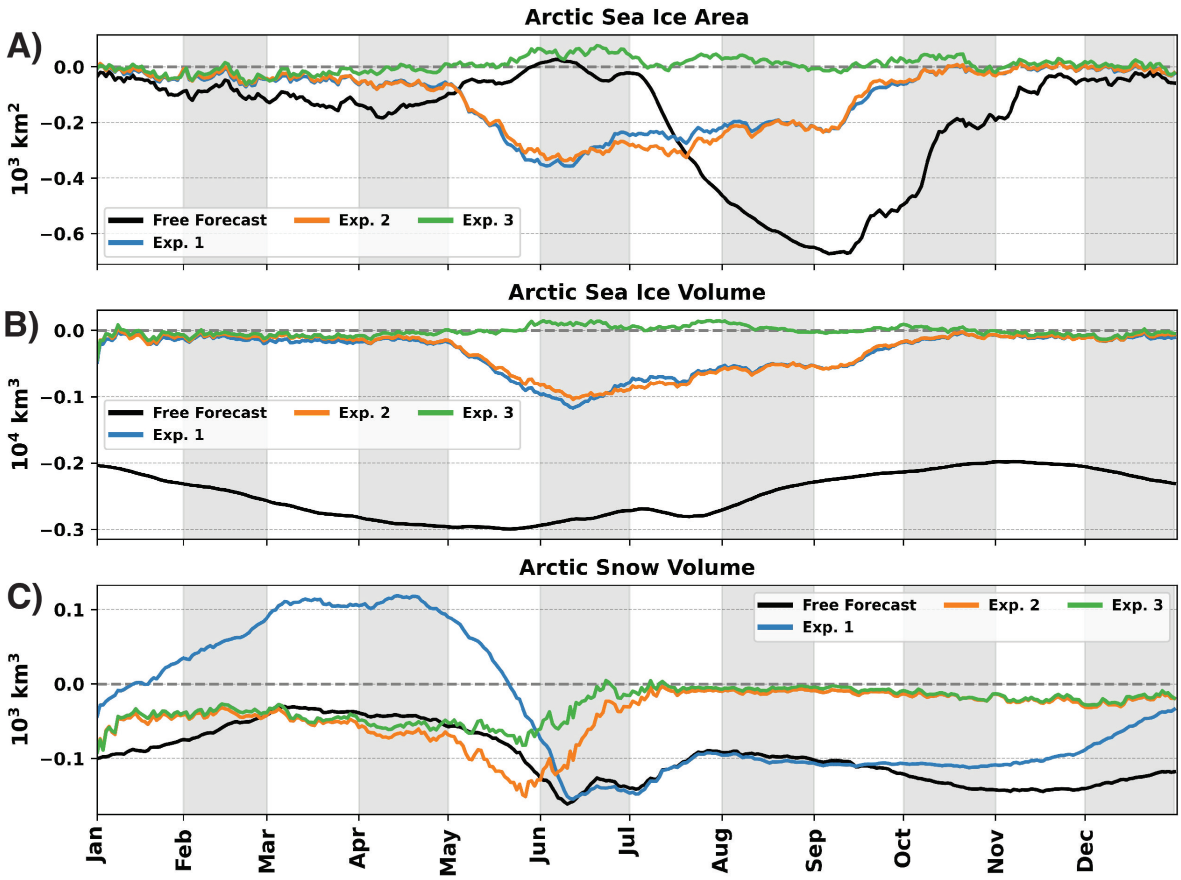

Fig. 1) Same as Figure 3 in the manuscript.

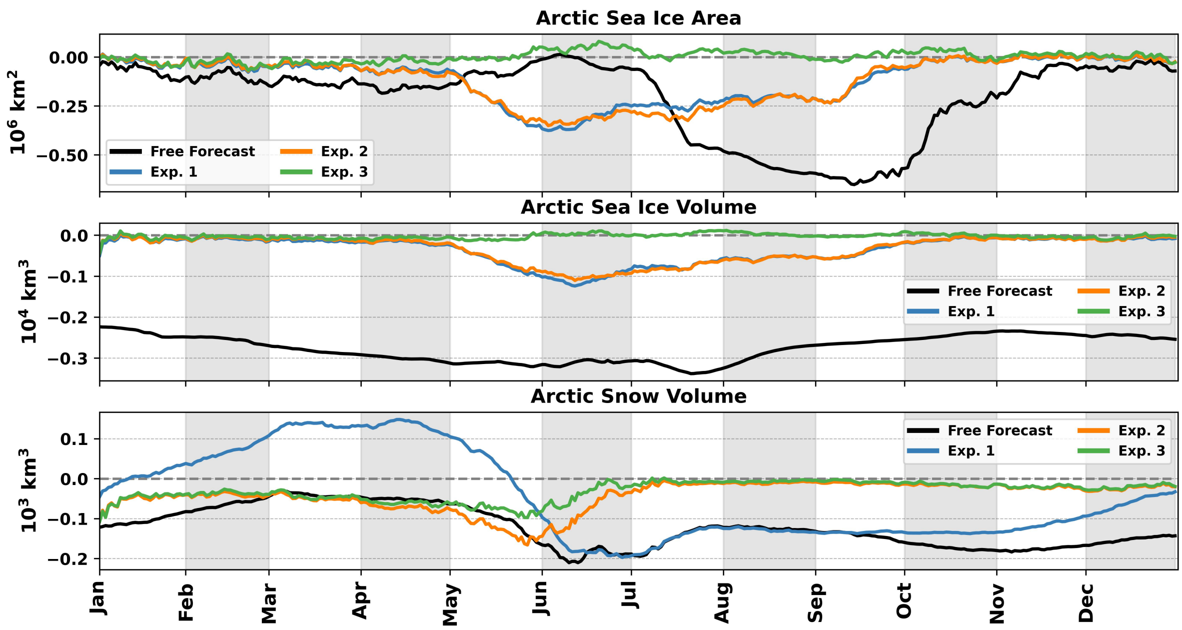

Fig. 2) Same as Figure 3 in the manuscript but using the ensemble median instead of the ensemble mean.

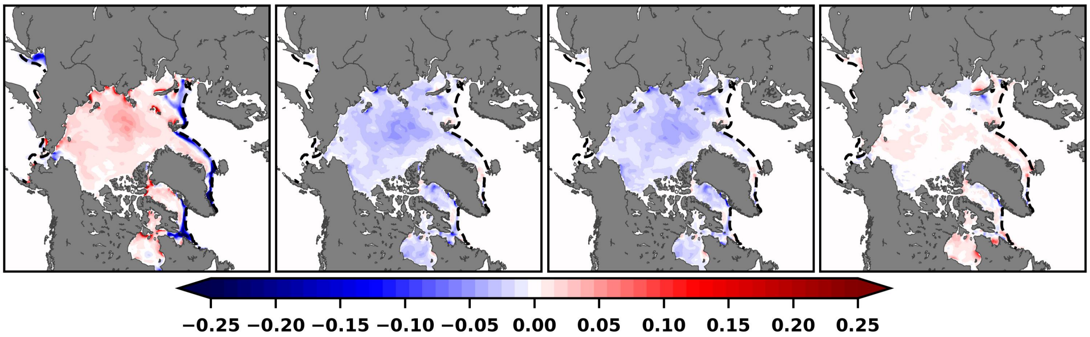

Fig. 3) Same as Figure 5 in the manuscript.

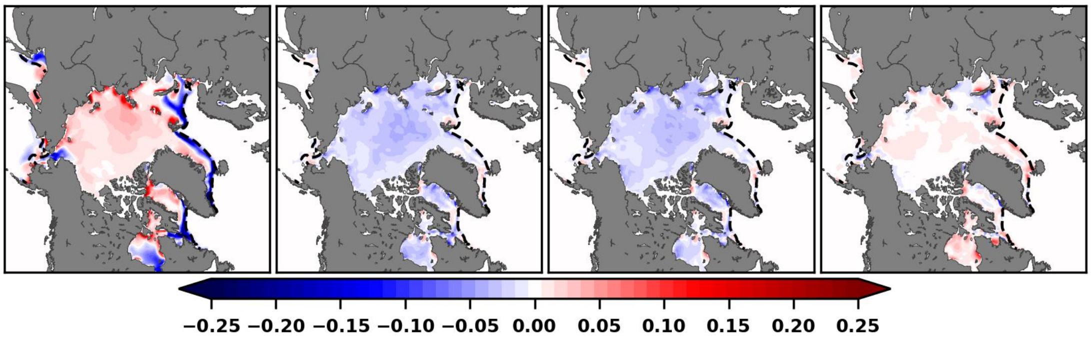

Fig. 4) Same as Figure 5 in the manuscript but using the ensemble median instead of the ensemble mean.

The previous paragraph concerning apparently biased distributions relates to a more general consideration of whether or not it makes any sense to evaluate a bias or using any other way of evaluating a single quantity, such as the mean, that has been extracted from the ensemble distribution. More general approaches, such as the continuous ranked probability score (CRPS), attempt to evaluate the accuracy of the entire ensemble distribution and not a single value obtained from the distribution. Please consider using such an approach for evaluating the resulting ensembles or explain why it hasn't been done.

Another reviewer mentioned adding an additional evaluation method that uses the ensemble instead of the mean. I chose to included the spatial probability score, which can be considered a spatial analogue of CRPS.

Line 65: "...if there is an optimal data assimilation setup...": This claim seems much too general, since the optimal configuration will likely depend on many details of a particular application of DA to sea ice. One such detail is the observing network, especially as mentioned elsewhere with respect to the unrealistic spatial distribution of SIC observations. Please rephrase this here and elsewhere.

This comment was addressed in the text.

line 65: "The OSSEs presented in this study will test different experimental setups to investigate their impacts on sea ice and snow states generated by data assimilation."

line 432: "These additional experiments would further help us understand the correct data assimilation setup for representing sea ice and snow in climate analyses."

Line 83: “One unique aspect of the EAKF. . . ”: This is misleading, since all variants of the ensemble Kalman filter use flow-dependent background-error covariances, not just the EAKF. Also, other data assimilation approaches, such as ensemble-variational approaches, use flow-dependent background-error covariances. Please rephrase.
This comment was addressed in the text.
line 83: "One important aspect of the EAKF is the ability to use a flow-dependent background-error covariance, which differs from a static background-error covariance typically employed by variational techniques."

Line 91: “. . . poor representation of model errors. . . ”: Presumably in an OSSE context you can perfectly represent any model errors that you choose to include in the experimental setup. Therefore, if inflation is still needed in such a context, then it must be serving a different purpose than accounting for model error. Please give some explanation for what these purposes are.
This was addressed in the text.
lines 90-94: "Adaptive prior covariance inflation was applied by “inflating" the prior background fields, increasing the variance by pushing ensemble members away from the ensemble mean (Anderson, 2007). Zhang et al. (2018) found a reduction in total Arctic sea ice area and volume errors when prior inflation was applied in their study."

Line 133: “. . . 15%of the true values of SIC. . . ”: It is really 15% of the true values of SIC? Meaning that open water and very low concentration values are perfectly observed (i.e. error standard deviation close to zero)? Or is the standard deviation simply 15% (and not dependent on the true SIC)?
I added some text to the manuscript to make this comment more clear in the text.
line 133: "The observation error standard deviation for SIC is 15% of the true values of SIC (SICerror = SICtruth*0.15; Zhang et al. (2018)) and 0.1 m for SIT (approximation of future high precision data; Zhang et al. (2018))."
The observation error standard deviation is set by multiplying the truth SIC value by 15% (SIC_{error} = SIC_{truth}*0.15). This method of specifying the SIC observation error follows the same method used in Zhang et al. (2018), which they state is "an approximate combination of bias and precision of the satellite-based concentration; e.g., Meier 2005." While our largest SIC uncertainty would be 15%, numerous other studies have employed other static or formula based methods for specifying the SIC observation uncertainty that would give similar values compared to our study (Tonboe and Nielsen, 2010; Mathiot et al., 2012; Zhang et al., 2021; Lee and Ham, 2022; Cheng et al., 2023). Additionally, this study was more focused on the data assimilation setup and not the observational side. However, the authors do agree that focusing on the observational setup (observation errors, observation density, etc) for data assimilation experiments should be addressed in future studies. Since other reviewers had comments on the SIC observation error specification, the authors tried to address some of the questions using the simplified data assimilation experiments that was presented in the paper (see section 3.3).

Line 134: "The locations for all synthetic observation types...": For all observation types? This is very unrealistic for SIC which is well observed almost everywhere by passive microwave satellite observations every day. CryoSat-2 is only used for ice thickness measurements. The retrieval process for obtaining ice thickness from these measurements also depends heavily on having accurate snow depth information, which is not well observed currently by any instruments and therefore the ice thickness measurements can have very high levels of uncertainty. I think that for this study to be relevant, a more realistic observing network must be used for all observation types (and also realistic values for observation uncertainties).

The authors addressed the SIC uncertainty in the comment above. The sea ice thickness uncertainty follows that used in Zhang et al. (2018), however, the authors realize that real sea ice thickness uncertainties can vary due to the different conditions. A sentence was included in the paper stating this.

line 135: "While studies that use real SIT observations have varied their uncertainties depending on their thickness value (Xie et al., 2018; Cheng et al., 2023), due to the complexity of computing SIT (Zygmuntowska et al., 2014) this study chose to use a single value for SIT uncertainty."

The authors did include a note in the text that our uncertainty value is for potential future datasets that might have higher precision. However, the sea ice thickness uncertainty chosen for this study is not too far off from values found in Zygmuntowska et al. (2014) (0.28 m in February/March and 0.21 m in October/November). In regards to the comment about the observing network, the authors chose the observing network for easy experimental setup and fair comparison between the observations that are assimilated in this study. A sentence was added in the text explaining why the network was chosen.

Line 136: The acronyms such as AICEN, VICEN, VSNON, SNWD, etc. are very non-intuitive and difficult to remember. These look like FORTRAN variable names, which are not necessarily appropriate for a scientific paper. Please consider variable names or non-acronym labels that readers will more easily remember and recognize. For example, if a quantity is a function of the thickness category, then this could be represented by a subscript of a variable corresponding to the thickness category index.

This comment has been addressed in the text. All references made to AICEN, VICEN, and VSNON have been replaced with category-based sea ice area, category-based sea ice volume, and category-based snow volume. Variables summed up over the different thickness categories are now referred to as SIC, V_{ice}, and V_{snow}. Observations are now referenced using SIC, SIT, D_{snow}, and SIST. I decided to leave the acronym for sea ice surface temperature (SIST) because it is close to the acronym for sea surface temperatures (SST) and easier to remember.

Line 226: "...not assimilating SIC observations improves most forecast metrics...": A more realistic (i.e. much denser) observing network for SIC would likely lead to more improvement when assimilating SIC since it would also lead to less ensemble spread and therefore reduced non-Gaussian effects. Also, as already mentioned, it's not clear if the observation error standard deviation for SIC observations is state dependent and, if so, if this could cause some of the resulting negative bias since low SIC observations will obtain more weight than high SIC observations.

As can be viewed below in Figures 5 and 6, the spread amongst the ensemble members for experiments 2 and 3 are quite small. This means the observations are having quite an impact on collapsing the initial spread that was found in the free forecasts (see Figure 1 in manuscript). The observation error specification was addressed in previous comments above. The authors believe that the underlying issue is really related to the boundedness of the observation error distribution (bounded above 1 and below 0 for SIC) and the formulation of the EAKF and RHF using a Gaussian observation likelihood. This mis-match will lead to bias solutions which other studies have shown (see text for citations to these studies). Since SIC is doubly bounded, generated observations would be inherently biased near either bound of zero or one. Our observations would be more biased near 1 in our study since their observation errors would be large based on our specification. Even with higher SIC observation errors occurring for SIC values near 1 (over the central Arctic), the observations still have enough weight to pull our ensemble towards being negatively biased. The authors tried to address some of the questions regarding the observation error specification impacts on this study using the simplified data assimilation experiments (section 3.3).

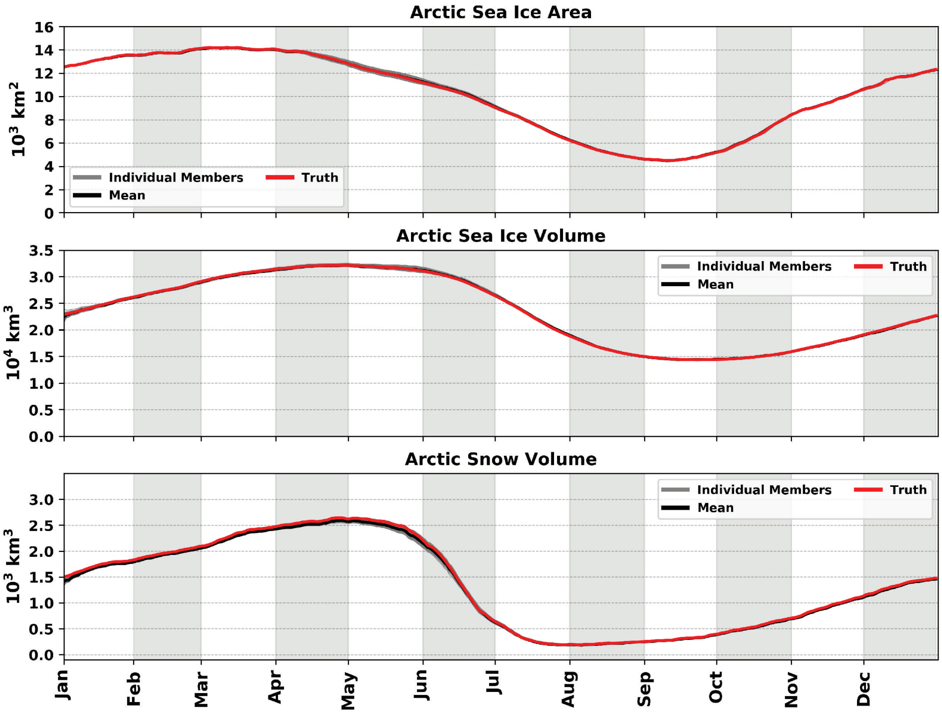

Fig. 5) Daily total Arctic sea ice area, sea ice volume, and snow volume from experiment 2 where SIC, SIT and D_{snow} observations are assimilated. Each gray line represents an individual ensemble member, black line represents the ensemble mean, and the red line represents the truth member.

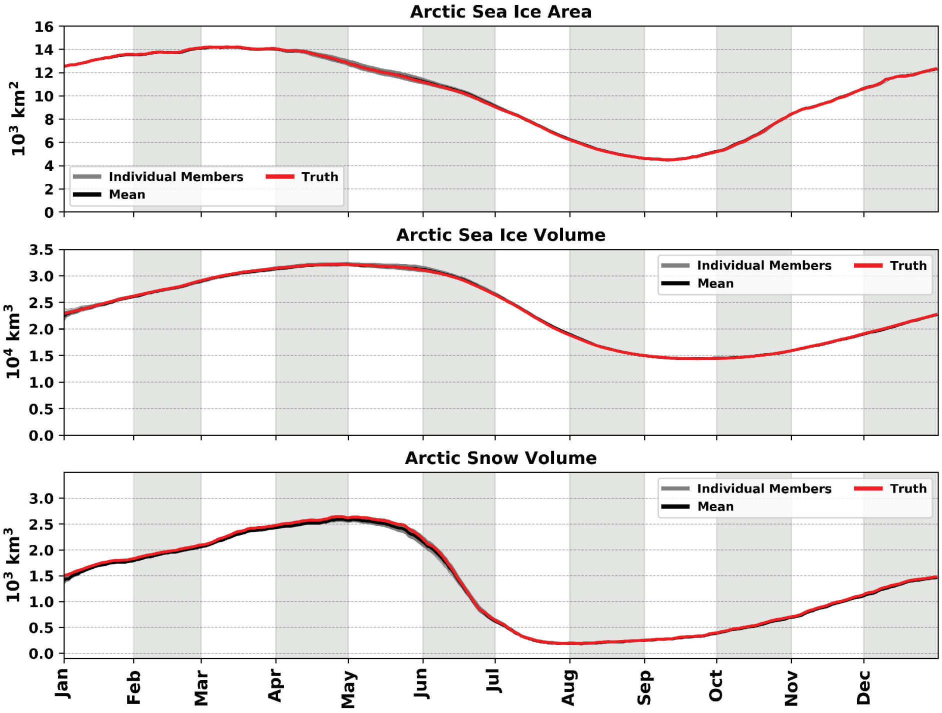

Fig. 6) Daily total Arctic sea ice area, sea ice volume, and snow volume from experiment 3 where SIT and D_{snow} observations are assimilated. Each gray line represents an individual ensemble member, black line represents the ensemble mean, and the red line represents the truth member.

References

Cheng, S., Chen, Y., Aydoğdu, A., Bertino, L., Carrassi, A., Rampal, P., and Jones, C. K.: Arctic sea ice data assimilation combining an ensemble Kalman filter with a novel Lagrangian sea ice model for the winter 2019–2020, The Cryosphere, 17, 1735–1754, 2023.

Lee, J.-G. and Ham, Y.-G.: Satellite-Based Data Assimilation System for the Initialization of Arctic Sea Ice Concentration and Thickness Using CICE5, Frontiers in Climate, 4, 797 733, 2022.

Mathiot, P., König Beatty, C., Fichefet, T., Goosse, H., Massonnet, F., and Vancoppenolle, M.: Better constraints on the sea-ice state using global sea-ice data assimilation, Geoscientific Model Development, 5, 1501–1515, 2012.

Meier, W. N.: Comparison of passive microwave ice concentration algorithm retrievals with AVHRR imagery in Arctic peripheral seas, IEEE Transactions on geoscience and remote sensing, 43, 1324–1337, 2005.

Tonboe, R. and Nielsen, E.: Global sea ice concentration reprocessing validation report, EUMETSAT OSI SAF Prod. Rep. OSI-409, version, 1, 18, 2010.

Zhang, Y.-F., Bitz, C. M., Anderson, J. L., Collins, N., Hendricks, J., Hoar, T., Raeder, K., and Massonnet, F.: Insights on sea ice data assimilation from perfect model observing system simulation experiments, J. Climate, 31, 5911–5926, 2018.

Zhang, Y.-F., Bushuk, M., Winton, M., Hurlin, B., Yang, X., Delworth, T., and Jia, L.: Assimilation of satellite-retrieved sea ice concentration and prospects for September predictions of Arctic sea ice, Journal of Climate, 34, 2107–2126, 2021.

Zygmuntowska, M., Rampal, P., Ivanova, N., and Smedsrud, L. H.: Uncertainties in Arctic sea ice thickness and volume: new estimates and implications for trends, The Cryosphere, 8, 705–720, 2014.