This is a very well-written and timely manuscript introducing MET-AICE v1.0, the first operational data-driven short-term sea ice prediction system for the European Arctic. I particularly appreciate that the evaluation is based on the most recent year of operational forecasts, which makes the results highly convincing and relevant for end-users. The manuscript is clear, the methodology is carefully described, and the system itself represents a significant step forward for operational sea ice forecasting. The authors convincingly show that MET-AICE provides skilful 10-day sea ice concentration forecasts at 5 km resolution, consistently outperforming both persistence and the Barents-2.5 km dynamical prediction system. Overall, I find the paper well-suited for publication after minor revisions. Below, I provide some suggestions that I believe could further strengthen the manuscript.

Comments and questions

Prediction scheme design

The use of separate models for each lead time is interesting. Could the authors elaborate on why this design was preferred over a more common autoregressive approach? It would strengthen the manuscript to show why this scheme provides better sea ice concentration forecasts than autoregression. Additionally, once the 10 forecasts are concatenated, do they retain physical consistency from one time step to the next?

Thanks for this comment. We agree that a discussion about that is missing in the preprint. We have added the following figure in the paper:

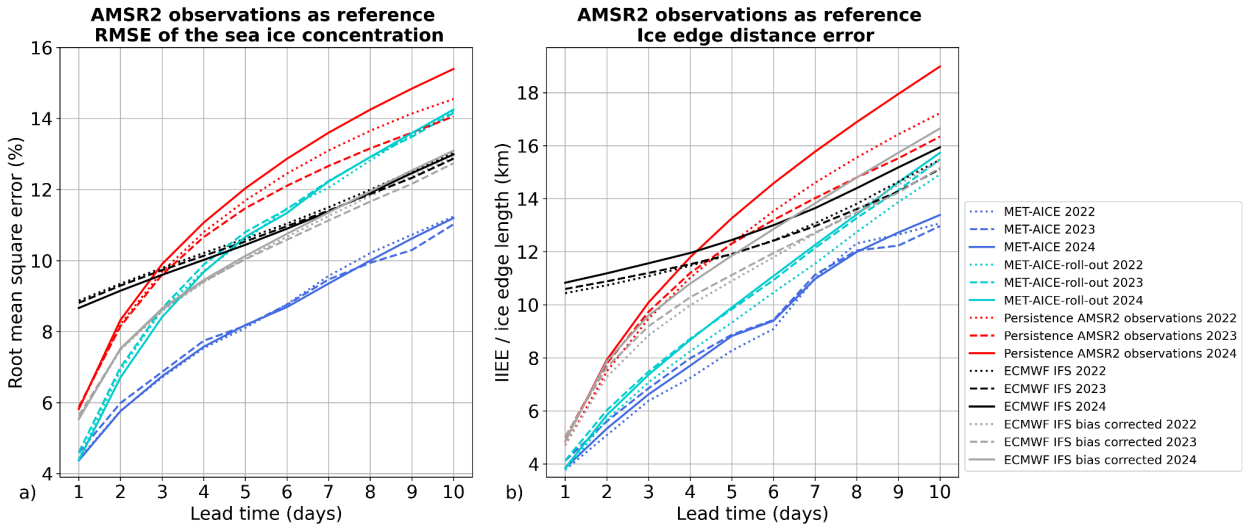

Figure 4. Performances of MET-AICE and ECMWF IFS during the years 2022, 2023, and 2024 over the full MET-AICE domain using AMSR2 observations as reference. MET-AICE-roll-out corresponds to using the MET-AICE model for 1-day lead time auto-regressively to predict the sea ice concentration for lead times up to 10 days. a) Root mean square error (RMSE) of the

sea ice concentration. b) Evaluation of the ice edge position (defined by the 10 % sea ice concentration contour).

And the following paragraph in the results section to discuss this:

"MET-AICE and ECMWF IFS forecasts are evaluated during the period 2022 - 2024 over the full domain of MET-AICE using AMSR2 observations as reference (Fig. 4). In addition, we evaluated the MET-AICE model developed for 1-day lead time in an auto-regressive mode to predict SIC up to 10 days ahead (hereafter referred to as "MET-AICE-roll-out"). In this configuration, the SIC prediction from the previous lead time is used as a predictor to predict the next time step. While this configuration allows to improve the consistency between the time steps, it results in poorer performances than the operational version of MET-AICE. On average over all lead times, the RMSE of the SIC is about 19 % larger for MET-AICE-roll-out than for the operational version of MET-AICE, and the error for the ice edge position is about 10 % larger for MET-AICE-roll-out. Therefore, we decided to only present the results for the operational version of MET-AICE for the rest of this study."

Forcing model evolution

At line 120, the manuscript notes the distribution shift in ECMWF atmospheric forecasts from one cycle to the next. To what extent do the authors envisage the need for retraining MET-AICE in the coming years as ECMWF forecasts continue to evolve? Have they evaluated model skill across past ECMWF cycles to quantify the impact?

We have added a new figure (see below) showing the interannual variability in the performances of MET-AICE and ECMWF IFS between 2022 and 2024. MET-AICE has very similar performances during these three years, suggesting that re-training the system is not necessary every year. We plan to re-train the models every time there will be modifications in the MET-AICE prediction system.

The following paragraph has been added:

"For all lead times and all the years evaluated, MET-AICE considerably outperforms persistence of AMSR2 observations (RMSE of the SIC and ice edge distance error about 28 % smaller on average), ECMWF IFS forecasts (RMSE of the SIC and ice edge distance error about 25 % and 30 % smaller, respectively), as well as ECMWF IFS bias corrected (RMSE of the SIC and ice edge distance error about 18 % and 21 % smaller, respectively). Furthermore, there is little interannual variability in the performances of MET-AICE, which suggests that re-training MET-AICE does not have to be done every year."

Furthermore, we also evaluated this in figure S3 of the supplement in Palerme et al., 2024 (https://tc.copernicus.org/articles/18/2161/2024/tc-18-2161-2024-supplement.pdf). Note that the same neural network architecture and almost the same data (but over a different domain) were used in Palerme et al., 2024. The conclusion of this analysis was that it was beneficial to use a

longer training period (and therefore data from more ECMWF model cycles) for the sea ice concentration forecasts.

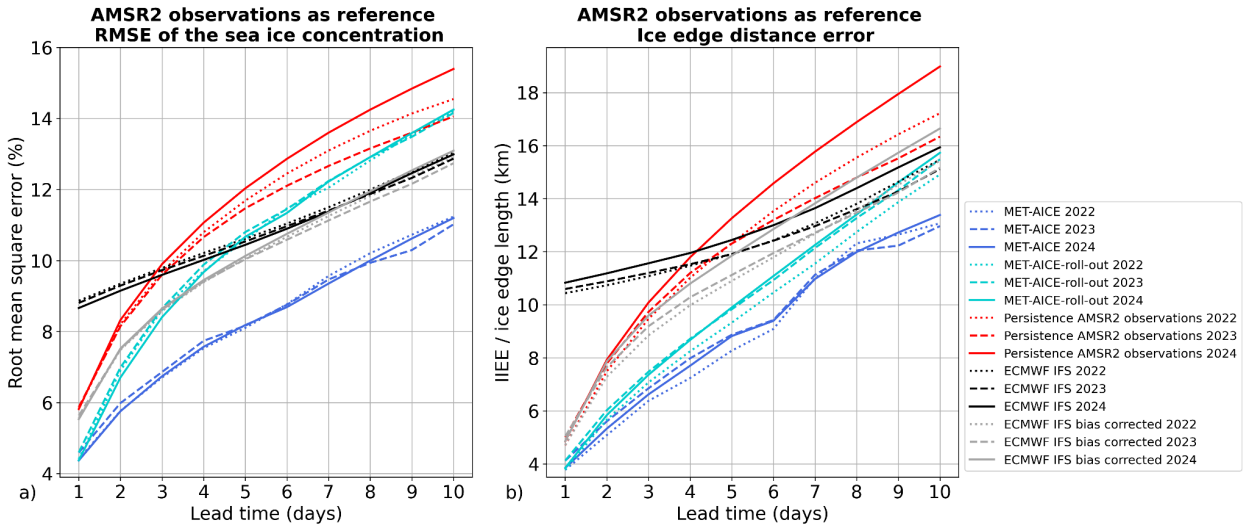

Figure 4. Performances of MET-AICE and ECMWF IFS during the years 2022, 2023, and 2024 over the full MET-AICE domain using AMSR2 observations as reference. MET-AICE-roll-out corresponds to using the MET-AICE model for 1-day lead time auto-regressively to predict the sea ice concentration for lead times up to 10 days. a) Root mean square error (RMSE) of the sea ice concentration. b) Evaluation of the ice edge position (defined by the 10 % sea ice concentration contour).

Verification metric robustness

How robust is the ice edge distance error to differences in smoothing across products? I would expect a sharper forecast to yield a longer ice edge, due to more small-scale information, compared to a smoother ML forecast. Clarification here would be valuable.

Thanks for this very relevant comment. We performed a sensitivity experiment to investigate this and we added a figure showing the results of this experiment in the paper. The figure and the text added in the revised version of the paper:

In the results section:

"In order to investigate to what extent the verification scores used in this study are influenced by the effective spatial resolution of the SIC fields, a sensitivity experiment was performed (Fig. 3). The AMSR2 SIC observations from 2013 to 2024 were used to calculate the RMSE of the SIC and the ice edge distance error for 1-day persistence of AMSR2 retrievals. The observations from the first day were smoothed using various Gaussian filters with a standard deviation ranging from 5 km (1 grid point) to 30 km (6 grid points), and compared to the original AMSR2 SIC field from the second day. Only the ice edge length from the observations of the second day at the original spatial resolution (5 km) was used to compute the ice edge distance error

regardless of the size of the Gaussian filter. In order to avoid some biases resulting from smoothing coastal grid points, the grid points closer than 50 km from the coastlines were removed from this analysis. Smoothing the AMSR2 SIC fields results in lower RMSE if the standard deviation of the Gaussian filter is between 5 and 20 km. The lowest RMSE is obtained for a Gaussian filter with a standard deviation of 10 km, with a reduction of 8.4 % compared to the RMSE obtained using the original SIC fields at 5 km resolution. This is in contrast with the ice edge distance error, which is only reduced by 0.5 % when a Gaussian filter with a standard deviation of 5 km is applied compared to the score obtained with the original resolution. Moreover, the ice edge distance error constantly increases when larger Gaussian filters are applied. Therefore, the ice edge distance error does not really favor smoother SIC fields contrary to the RMSE of the SIC."

In the conclusion:

"Furthermore, it is common practice to evaluate machine learning models using verification scores that are strongly correlated with the loss function. This study highlights that this can lead to spurious conclusions if no independent verification score is used in addition. Here, we show that the RMSE of the SIC can be reduced by 8.4 % due to the smoothing of the SIC fields. Therefore, we strongly recommend using more independent verification scores, such as the ice edge distance error, which does not favor smoother SIC fields as the RMSE of the SIC does."

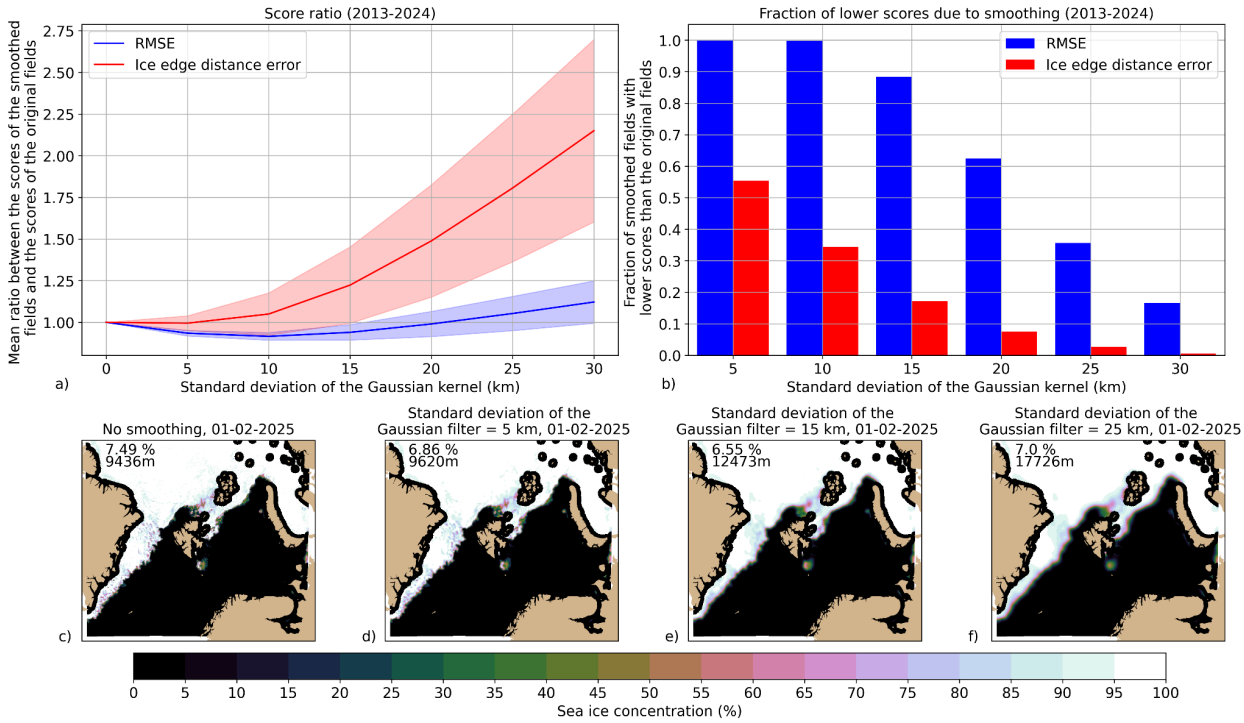

Figure 3. Influence of the effective spatial resolution on the verification scores. The AMSR2 observations during the period 2013 - 2024 were used to calculate the RMSE of the sea ice concentration and the ice edge distance error for 1-day persistence of AMSR2 observations over the MET-AICE domain (4327 days evaluated). The AMSR2 observations from the first day were smoothed using various Gaussian filters with a standard deviation ranging from 5 km (1 grid point) to 30 km (6 grid points), while the original AMSR2 observations from the next day were used as reference for computing the scores. The ice edge distance error was computed using the ice edge length from the observations of the second day at the original spatial resolution (5 km). The grid points closer than 50 km from the coastlines were excluded from this analysis. a) The mean ratio between the scores after smoothing and the scores using the original data (a score lower than 1 means that the smoothing results in a better score). The shaded areas represent the standard deviations of the verification scores. b) The fraction of scores that are better after smoothing the AMSR2 sea ice concentration fields. c) AMSR2 sea ice concentration observations at the original spatial resolution (5 km) on 01/02/2025. d, e, f) AMSR2 sea ice concentration observations on 01/02/2025 smoothed using Gaussian filters with a standard deviation of 5 km (d), 15 km (e), and 25 km (f). The scores on the top left corners of the maps show the RMSE of the sea ice concentration (top, in %) and the ice edge distance error (bottom, in meters) for the comparison between the AMSR2 observations from 01/02/2025 and 02/02/2025

Sensitivity to different forcings

Have the authors tested MET-AICE under different atmospheric forcings than those used in training (e.g., AROME-Arctic)? Would the model's forecast skill improve or deteriorate in such cases, as seen in Barents-2.5?

We have not tested running MET-AICE under different atmospheric forcings, but it is likely that the forecast skill would deteriorate if another forcing (such as AROME-Arctic) would be used. A machine learning model is trained to learn the correlation (non linear) between features of a dataset (the predictors) and the target variables. When a dataset with different properties than the one used for training is used as a predictor, some biases can be expected in the predictions. For example, AROME-Arctic has a much higher spatial resolution (2.5 km) than ECMWF IFS (9 km). Since MET-AICE has been trained using ECMWF IFS forecasts, MET-AICE has learnt some convolutional kernels (3 x 3 grid points) at maximum 9 km resolution, and should not be able to take advantage of the higher spatial resolution of AROME-Arctic forecasts. It is also worth noting that AROME-Arctic does not cover the full domain of MET-AICE, and that convolutional neural networks are not designed to work on a grid with a varying resolution. For grids with varying resolutions, graph neural networks would be much more relevant than convolutional neural networks. Therefore, we did not include any experiment like this in the revised version of the paper.

Relatedly, it would be instructive to assess MET-AICE in controlled "idealised" experiments not present in the training data (e.g., constant zonal winds, uniformly melting temperatures across the domain). Do the forecasts behave in line with expected sea ice physics?

We have investigated this using either a constant wind field of about 10 meter per second with a north west direction (blowing towards the south east) or a constant 2-meter temperature field of 280 kelvins. The forecasts were evaluated during the period April 2024 to March 2025. On average, the forecasts with a constant wind field produced an increase of the sea ice extent as expected, and the forecasts with a constant temperature field of 280 K produced a decrease of the sea ice extent as expected. However, these results were not systematic. In particular for the constant wind field, an increase of the sea ice extent was observed in only 65 % of the cases, and there was a lot of spatial variability in these results. These results are difficult to interpret because of this variability and because of the lack of ground truth.

Furthermore, convolutional neural networks are trained to recognize spatial patterns through the convolution operations. On constant fields, the neural networks would not recognize any spatial patterns learnt during training, which would likely result in inaccurate predictions. Moreover, providing a dataset to a supervised neural network which is very different from any sample used for training the neural network would also likely result in inaccurate predictions. Therefore, we did not include any experiment like this in the revised version of the paper.

Along these lines, has the system been tested using ECMWF ensemble members in addition to the control forecast? Does MET-AICE produce a reasonable spread in SIC forecasts under such perturbations?

Thanks for this comment. We have investigated this in the revised version of the paper. Please find below the figure and the text added in the revised version of the paper:

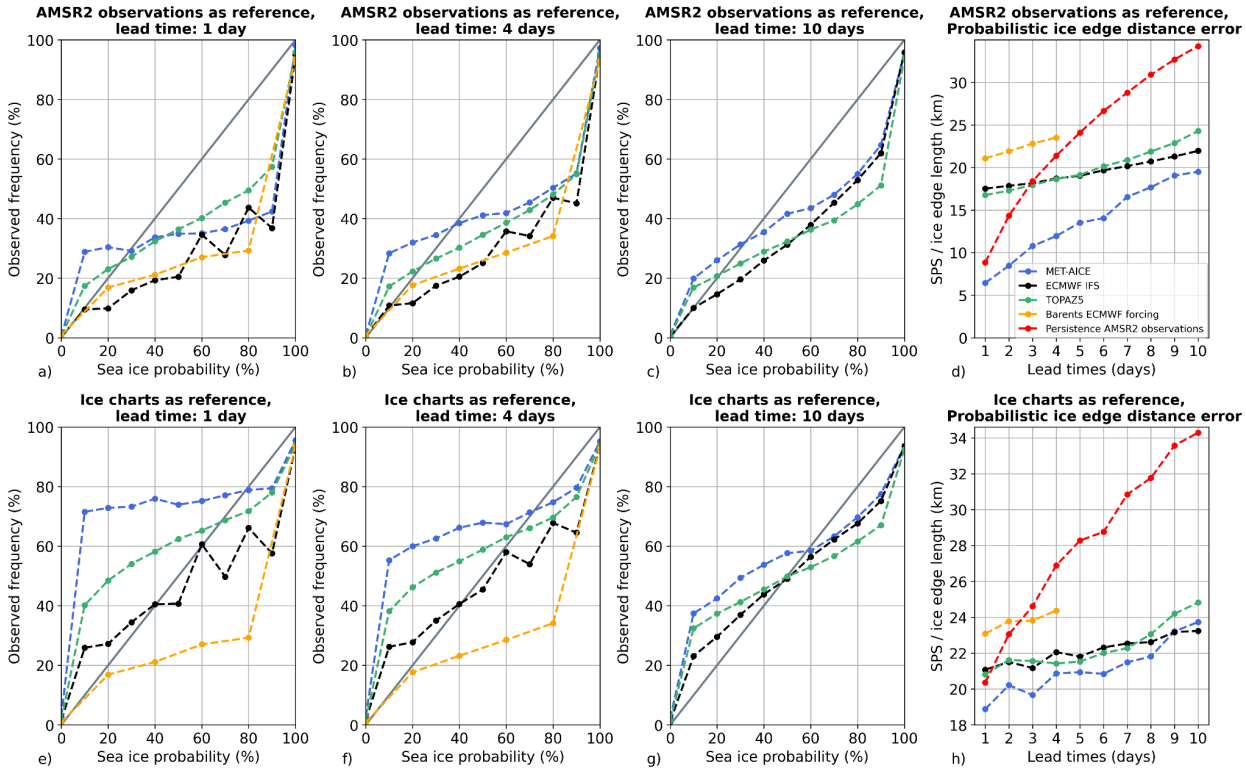

Figure 10. Evaluation of probabilistic sea ice forecasts. The forecasts starting on Mondays, Wednesdays, and Fridays from April 2024 to March 2025 are evaluated in the shared domain between MET-AICE and Barents-2.5. a, b, c) Comparison between the sea ice probability (probability that the sea ice concentration exceeds 10 %) and the observed frequency of occurrence in AMSR2 observations. d) Probabilistic ice edge distance error (ratio of the Spatial Probability Score over the observed ice edge length) depending on the lead time using AMSR2 observations as reference. e, f, g) Comparison between the sea ice probability and the observed frequency of occurrence in the ice charts. h) Probabilistic ice edge distance error depending on the lead time using the ice charts as reference.

In the Dynamical models section (section 2.2.2):

"This study primarily focuses on evaluating the MET-AICE operational forecasts that are deterministic and have daily time steps. Therefore, the daily means of SIC fields from the dynamical models are evaluated. These datasets consist of the high-resolution ECMWF IFS (HRES) forecasts, the TOPAZ5 ensemble mean, the Barents-2.5 member forced by AROME-Arctic, and the mean of Barents-2.5 members forced by ECMWF-ENS forecasts. In

addition, we also investigated producing probabilistic forecasts with MET-AICE in this study. For this experiment, MET-AICE forecasts were compared to the first 10 ensemble members from ECMWF IFS (ENS), the 10 TOPAZ5 ensemble members, and the 5 Barents-2.5 ensemble members forced by ECMWF-ENS produced at 00:00 UTC."

In the verification section (section 3.2):

"While deterministic SIC forecasts are evaluated in most of this study, an experiment was performed with probabilistic forecasts. The ice edge position in probabilistic forecasts is assessed using the sea ice probability ($SIP_{forecasts}$ in equation 2), which is defined as the fraction of ensemble members with a SIC higher or equal to 10 %, whereas a binary sea ice probability is used for the observations ($SIP_{observations}$ in equation 2). Then, the forecasts are evaluated using the ratio of the Spatial Probability Score (SPS, Goessling and Jung, 2018) over the ice edge length in the observations used as reference (hereafter referred to as "probabilistic ice edge distance error", equation 2). This metric was introduced by Palerme et al. (2019) and can be considered as the probabilistic extension of the ice edge distance error."

In the results section:

"In addition to evaluating the operational MET-AICE deterministic forecasts, we investigated if we could use several ensemble members from ECMWF IFS (ECMWF-ENS) for producing probabilistic SIC forecasts with MET-AICE. We used the first 10 ensemble members of ECMWF-ENS for the atmospheric predictors (10-meter wind and 2-meter temperature) to produce a set of 10 ensemble members with MET-AICE. In Fig. 10, the forecasts starting on Mondays, Wednesdays, and Fridays from April 2024 to March 2025 are evaluated in the shared domain between MET-AICE and Barents-2.5 in order to include the three dynamical models. Overall, MET-AICE and the three dynamical models produce ensemble forecasts that are overconfident (not enough ensemble spread), which means that low sea ice probabilities (SIP) are observed more frequently than predicted whereas high SIP are observed less frequently than predicted. While the MET-AICE forecasts are particularly overconfident for 1-day lead time, they have a similar spread as the TOPAZ5 forecasts for 10-day lead time. Among the prediction systems, ECMWF IFS produces the largest ensemble spread. Furthermore, the probabilistic ice edge distance error is lower in MET-AICE forecasts than in all dynamical models, except for lead times of 9 and 10 days when the ice charts are used as reference."

Missing Literature

I recommend citing also these papers: https://doi.org/10.3389/fmars.2023.1260047 and https://doi.org/10.1029/2024JH000433. Even though the latter does not pertain to the Arctic, it's still a good example of a data-driven prediction system targeting sea ice.

We have included these two references in the following sentence of the introduction:

“Several studies have recently shown that data-driven sea ice forecasting systems trained on satellite observations can be skillful (e.g. Grigoryev et al., 2022; Ren et al., 2022; Chen et al., 2023; Keller et al., 2023; Lin et al., 2023; Koo and Rahnemoonfar, 2024), and can provide more accurate forecasts than dynamical models (Andersson et al., 2021; Palerme et al., 2024; Kvanum et al., 2025; Lin et al., 2025). “

Citation: https://doi.org/10.5194/egusphere-2025-2001-RC2