Review of " MET-AICE v1.0: an operational data-driven sea ice prediction system for the European Arctic" by Palerme et al. Submitted to Geoscientific Model Development.

General comments

MET-AICE v1.0 is the first operational, data-driven sea ice prediction system specifically designed for short-term forecasts (1-10 days) in the European Arctic. The system is optimised for operational utility and higher spatial resolution, making it suitable for day-to-day maritime applications. The development of the MET-AICE system is particularly timely given the increasing demand for reliable, short-term, high-resolution sea ice forecasts, driven by increased maritime activity and heightened navigational risks associated with changing sea ice cover.

MET-AICE was trained on weekly AMSR2 weekly sea ice concentration data at 5-km resolution 2020 from the recently published reSICCI3LF algorithm, covering the period from 2013 to 2020. During training, the neural network models were iteratively updated over 100 epochs to minimize the mean squared error between the predicted SIC and the AMSR2 SIC observations. The system incorporates several predictors, including 9-km resolution ECMWF weather forecasts (2-m temperature and 10-m wind components), AMSR2 SIC observations from the day preceding the forecast start date, and a land-sea mask. MET-AICE uses a convolutional neural network with a U-Net architecture, designed specifically to capture spatial hierarchies in the input data. Operational forecasts have been generated since March 2024, with validation described in the manuscript covering a year-long period from April 2024 to March 2025. Despite demonstrated strengths in computational efficiency and accuracy compared to the Barents-2.5 km EPS model and other validation datasets, MET-AICE experiences reduced accuracy in coastal regions and diminished predictive skill during sea ice minimum periods, primarily related to inherent limitations in the input datasets. The current version of MET-AICE provides deterministic forecasts of sea ice concentration, which become smoother as the lead time increases. In future iterations, the authors plan to incorporate ensemble and probabilistic approaches to better quantify and represent the forecast uncertainty.

The paper is generally well written and structured, providing an important contribution towards operational high-resolution sea ice forecasting. However, several points need clarification before I can recommend the manuscript for publication.

I found the model description quite hard to follow. I wonder if you could include a flow diagram that shows the data inputs and preprocessing steps, a high level overview of the model architecture and key features (residual connections, spatial attention block and their purpose; downsampling and upsampling operations and progression of convolution kernels), and the outputs.

Thank you for this comment. We have added the flow diagram shown below in the revised version of the paper. Note that this figure is shown on an entire page in the revised version of

the paper. This describes what you mentioned. We agree that it helps to understand the architecture used in MET-AICE.

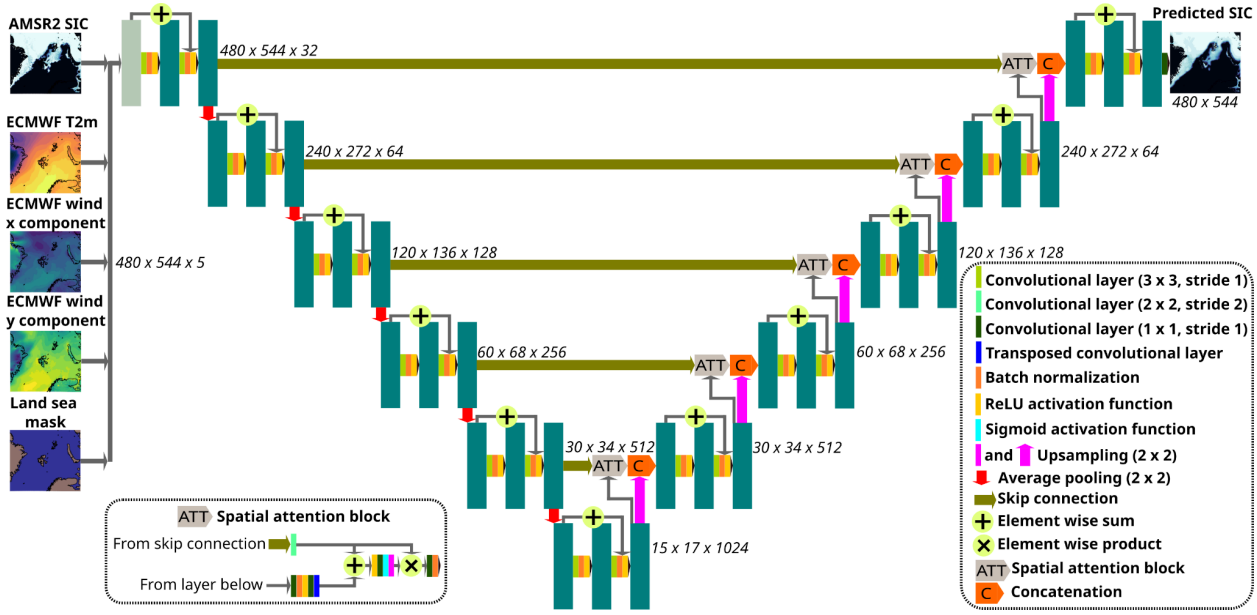

Figure 1. Architecture used in MET-AICE. The big green rectangles represent multi-channel feature maps. The dimensions of the feature maps (y, x, number of channels) are written next to each convolutional block.

The training period covers 7 years. Given the ongoing thinning and decline of sea ice cover, do you foresee a need for periodic retraining of the model? How might evolving sea ice conditions in the changing Arctic impact the model's forecasting accuracy over time?

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. We have added the following sentence in the results section:

"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."

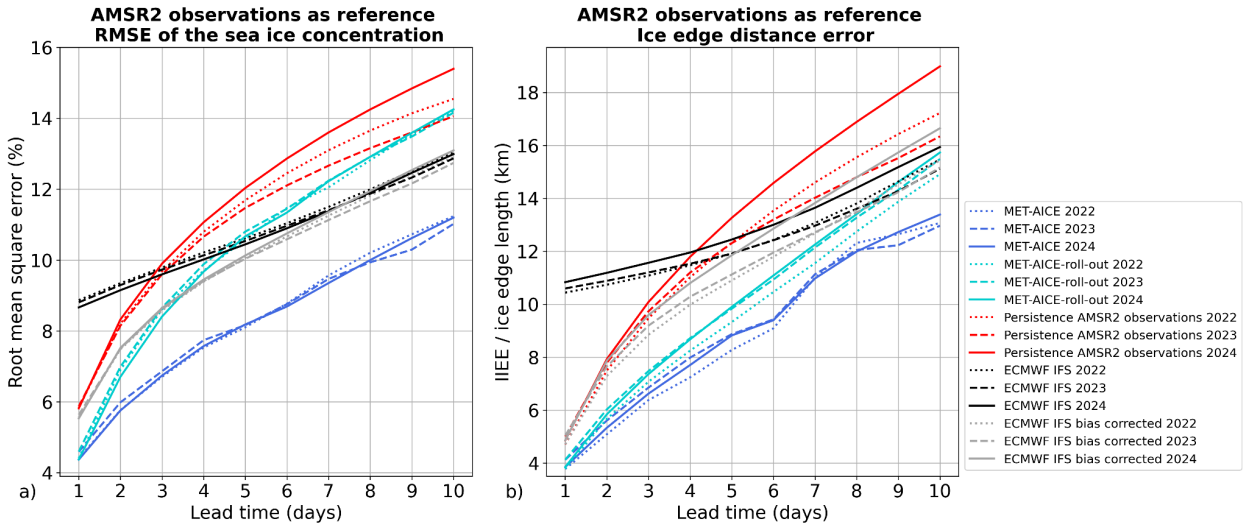

Figure 4. Performances of MET-AICE and ECMWF IFS during the years 2022, 2023, and 2024 over the full MET-AICE domain using AMSR2 observations as reference. MET-AICE-roll-out corresponds to using the MET-AICE model for 1-day lead time auto-regressively to predict the sea ice concentration for lead times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

Training is based on weekly datasets, yet the forecasts are daily. I presume that using weekly training data enhances the model's generalization capability and robustness against short-term noise? However, this choice may limit the model's ability to capture rapid, short-term sea ice dynamics occurring at daily scales. How does this choice impact forecast accuracy during periods of rapid sea ice changes? Is the reduced forecast skill during sea ice minimum periods possibly related to a temporal limitation inherent in weekly training data?

MET-AICE is trained on daily AMSR2 sea ice concentration observations. Since this was not clear in the preprint, we modified the following sentence:

P2, line 50 *"MET-AICE has been trained to predict SIC observations at about 5 km resolution derived from AMSR2 data using a three-step algorithm called reSICCI3LF (Rusin et al., 2024)."*

by:

"MET-AICE has been trained to predict daily SIC observations at about 5 km resolution derived from AMSR2 data using a three-step algorithm called reSICCI3LF (Rusin et al., 2024)."

And the following sentence:

P5, line 114 *"The deep learning models were trained using weekly data during the period 2013 - 2020 (about 400 forecasts for each lead time)"*

was replaced by:

"The deep learning models were trained using one forecast per week during the period 2013 - 2020 (about 400 forecasts for each lead time)."

The evaluation spans a single year of operational forecasts. Although this period enables an analysis of seasonal performance and highlights the reduced skill during the summer, significant year-to-year variability in sea ice conditions may affect the robustness of the conclusions drawn. How confident are you in your findings after just one seasonal cycle, and could interannual variability impact where and when the model performs well? I am mostly thinking of how you might ultimately assign an uncertainty flag to the forecast data product.

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.

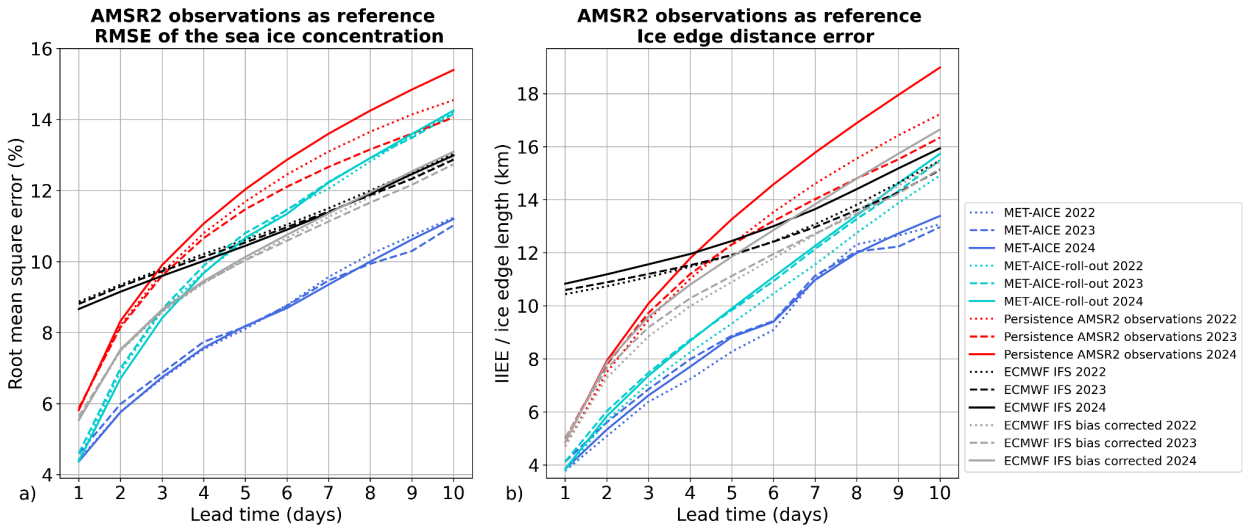

Figure 4. Performances of MET-AICE and ECMWF IFS during the years 2022, 2023, and 2024 over the full MET-AICE domain using AMSR2 observations as reference. MET-AICE-roll-out corresponds to using the MET-AICE model for 1-day lead time auto-regressively to predict the sea ice concentration for lead times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

Furthermore we have added the following paragraph to describe this figure:

"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. 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. While ECMWF IFS does not outperform persistence of AMSR2 observations for lead times up to 3 days, the bias correction allows a considerable improvement of the forecast skill for lead times up to 6 days"

The authors compare MET-AICE primarily to a single dynamical model, the Barents-2.5 km EPS. How does the performance of this dynamical model compare to other available dynamical models?

We have added a comparison with the dynamical models TOPAZ5 distributed by the Copernicus Marine Service and the ECMWF Integrated Forecasting System in the revised version of the paper in order to provide more context. As an example, the figure below (which is in the revised version of the paper) shows the performances of MET-AICE compared to Barents-2.5km, ECMWF IFS, and TOPAZ5.

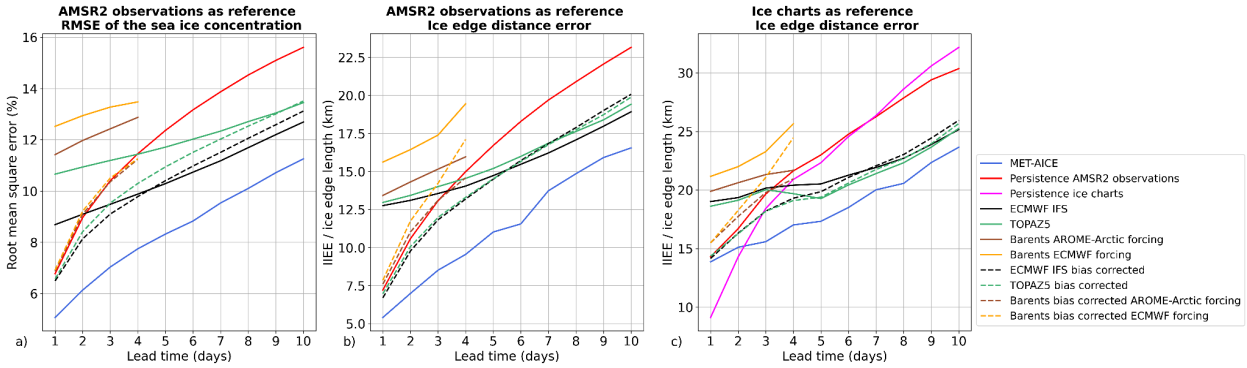

Figure 5. Performances of MET-AICE, ECMWF IFS, TOPAZ5 and Barents-2.5 during the period April 2024 - March 2025 over the shared domain between MET-AICE and Barents-2.5. a) Root mean square error (RMSE) of the sea ice concentration using AMSR2 observations as reference. b) Evaluation of the ice edge position (defined by the 10 % sea ice concentration contour) using AMSR2 observations as reference. c) Evaluation of the ice edge position using the ice charts as reference.

Specific comments

Line 63: It seems sensible to use 2-m temperature and 10-m winds to drive the system and you mention in the introduction that sea ice changes on short-time scales are driven by the wind. But was there any assessment of the optimal variables to train and run the model? At the very least it would be helpful to include references to justify your use of these variables to drive sea ice variability.

We think that this choice was already justified in the introduction of the preprint in lines 36 to 41:

"Sea ice changes on short-time scales are primarily driven by the atmosphere, and in particular by the wind (Mohammadi-Aragh et al., 2018; Yu et al., 2020). Hence, it is crucial to include predictors from weather forecasts when developing data-driven sea ice prediction systems, as suggested by previous studies. Grigoryev et al. (2022) reported an improvement of 5 to 15 % when forecasts from the National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS) are used, whereas Palerme et al. (2024) assessed an error reduction of 7.7 % when using ECMWF weather forecasts in addition to sea ice predictors."

Nevertheless, we modified this paragraph with adding details about the atmospheric variables used in the studies from Grigoryev et al. (2022) and Palerme et al., 2024. The new paragraph is:

"Sea ice changes on short-time scales are primarily driven by the atmosphere, and in particular by the wind (Mohammadi-Aragh et al., 2018; Yu et al., 2020). Hence, it is crucial to include predictors from weather forecasts when developing data-driven sea ice prediction systems, as suggested by previous studies. Grigoryev et al. (2022) reported an improvement of 5 to 15 % when using forecasts of 2-meter temperature, surface pressure and wind from the National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS). Moreover, Palerme et al. (2024) assessed an error reduction of 7.7 % when using ECMWF forecasts of 2-meter temperature and 10-meter wind in addition to sea ice predictors."

Line 65: I don't understand how the 10 different models were developed. Are each of these models for the different lead times, i.e. a set of 10 distinct forecasts for lead times of 1 day, 2 days, 3 days, all the way up to 10 days? Could you clarify the description here? Also, why do you have these different lead times - was the aim to find an appropriate lead time? Which is the dataset released via THREDDS? Is this the daily forecast with a 10-day lead time?

Yes, you are right. We developed a set of 10 distinct forecasts for lead times from 1 to 10 days. In the revised version of the paper, we added a comparison with a roll-out approach, in which the model for 1-day lead time is used to predict longer lead times auto-regressively (figure below). In this configuration, the sea ice concentration prediction from the previous lead time is used as a predictor to predict the next time step.The following sentences have been added to describe this comparison:

"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, this 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."

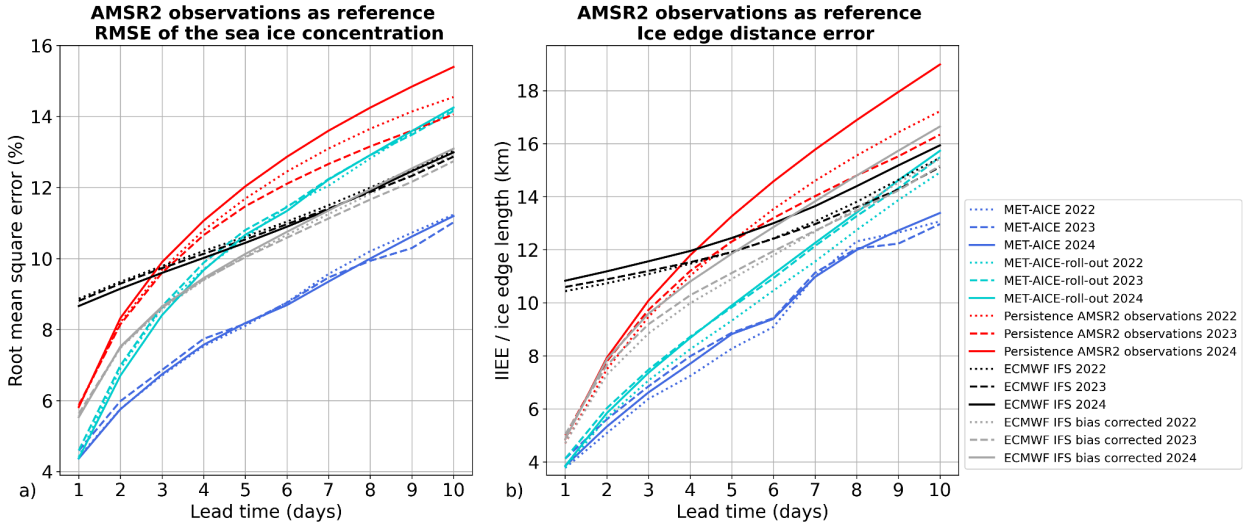

Figure 4. Performances of MET-AICE and ECMWF IFS during the years 2022, 2023, and 2024 over the full MET-AICE domain using AMSR2 observations as reference. MET-AICE-roll-out corresponds to using the MET-AICE model for 1-day lead time auto-regressively to predict the sea ice concentration for lead times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

Line 74-75: Coastal grid points (within 20 km of the coast) are excluded from the model performance evaluation. I didn't notice these points being masked out or flagged in some way in the forecasts released via the THREDDS server of the Norwegian Meteorological Institute. Might it be helpful to users if there is an indication of where you have confidence in the available forecast data and where users should take care.

Coastal grid points (within 20 km from the coast) are only excluded when the forecasts are evaluated using AMSR2 observations as reference, but not when the ice charts are used as reference because the ice charts are primarily based on higher-resolution satellite observations. In order to clarify this point, we have added the following sentence in section 2.2.1 describing the observations used in this study:

“Therefore, land contamination is much less present in the ice charts than in passive microwave observations, and we decided to take into account all oceanic grid points when the ice charts are used as reference (no coastal grid points excluded).”

Furthermore, while land contamination can be present in passive microwave observations and in the MET-AICE forecasts, it is not always the case. We decided to keep the coastal grid points in the forecasts delivered on the THREDDS server of the Norwegian Meteorological Institute, similarly to what is usually done for passive microwave sea ice concentration products. We do not provide any uncertainty estimates yet in the MET-AICE forecasts, but we decided to describe this issue in the paper instead.

Line 117: It isn't particularly clear how you used the datasets from 2021-2023 and why you only produced the validation on the data from April 2024 onwards. Would having a few extra years of validation assessment have made the results more robust?

We have added a new figure (see figure 4 above) 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. Furthermore we have added the following paragraph to describe this figure:

“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. 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. While ECMWF IFS does not outperform persistence of AMSR2 observations for lead times up to 3 days, the bias correction allows a considerable improvement of the forecast skill for lead times up to 6 days”

Technical corrections

Line 22: change “predict” to “predicts”

We suppose that you refer to line 222 here. We replaced “predict” by “predicts” in line 222.

Line 203: I think “less than” should be “fewer than” in this case

“less than” has been replaced by “fewer than”