We would like to thank the reviewers for their comments which helped us to improve the quality of the manuscript. Please find below our responses to the reviewer's comments.

Reviewer 2

##########

Review of Calibration of sea ice drift forecasts using random forest algorithms.

The manuscript describes a new method that post-processes numerical forecasts of sea ice drift using either in situ drifting buoys or satellite images for the training of a random forest algorithm. The results are evaluated against ice drift observations but in a different period, posterior to the training data. The results reveal that there is a systematic component of the ice drift forecast error that can be corrected by machine learning, although the reduction of error remains often less than 10%. The ML algorithms learns more efficiently from the buoys data than from the satellite images, highlighting the problem of temporal averaging.

The drift direction can mostly be improved in the short forecast range, likely because of the unpredictability of wind directions, but interestingly the algorithm is more often able to correct drift speed at longer forecast horizons, which I did not expect. The authors could spice up their article by analysing what their algorithm does to the sea ice drift speed that improves the skills at a 10 days range: are the drifts made systematically faster or slower? This kind of analysis can - if understood - lead to improvements of the forecast systems. More generally, not seeing what the algorithm does to the forecast is a little frustrating. An example of comparison of original to postprocessed and to observed sea ice drifts could be more convincing than cold-blooded skills scores.

We have added the following example of vector maps from TOPAZ4 forecasts and the calibrated forecasts in the supplementary material:

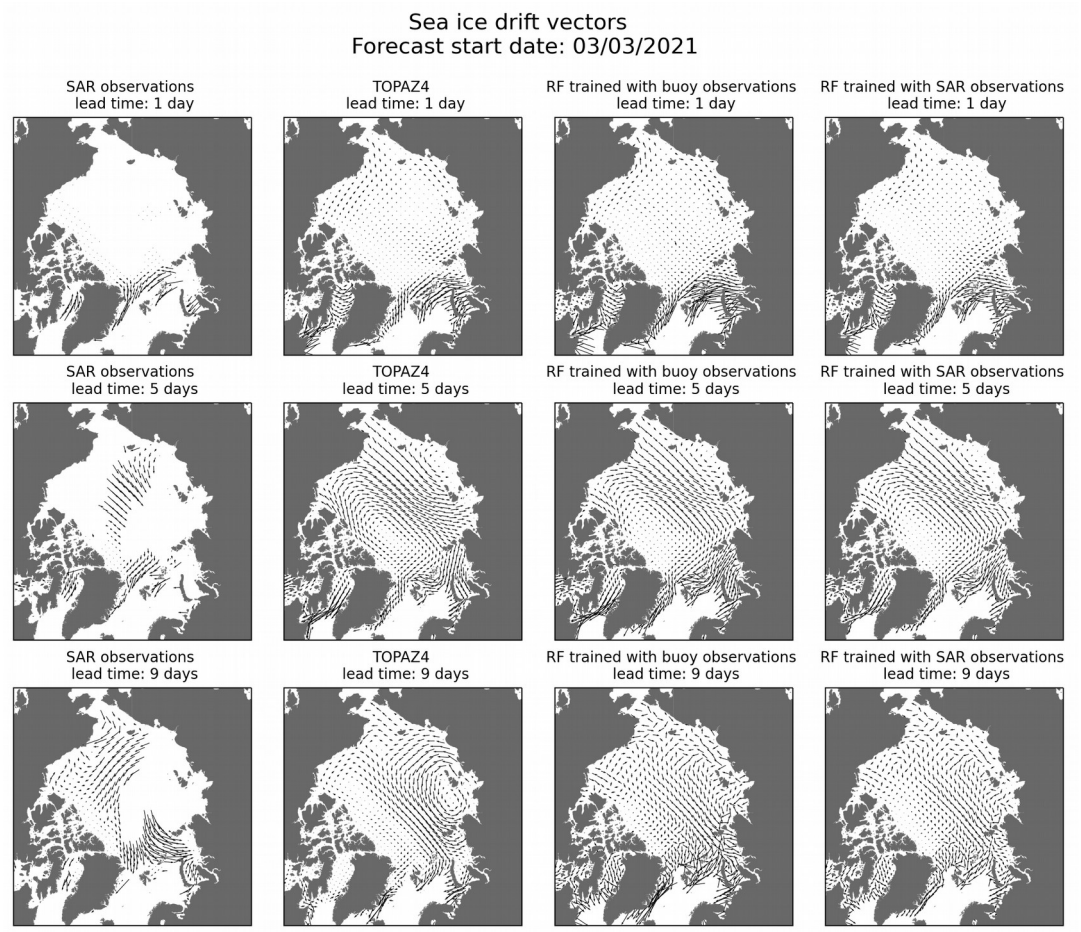


Figure S8. Example of calibration with the random forest (RF) algorithms for the forecasts which started on 03/03/2021 and for lead times of 1, 5, and 9 days.

Furthermore, we have also added the following figures showing the difference between the TOPAZ4 forecasts and the calibrated forecasts in the supplementary material:

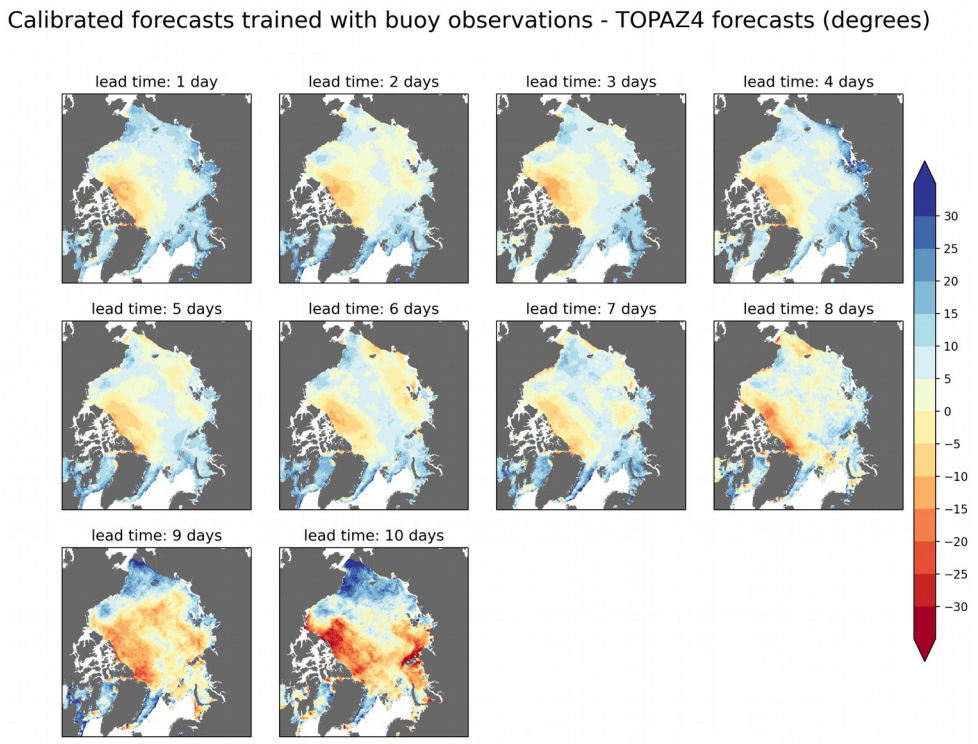


Figure S13. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

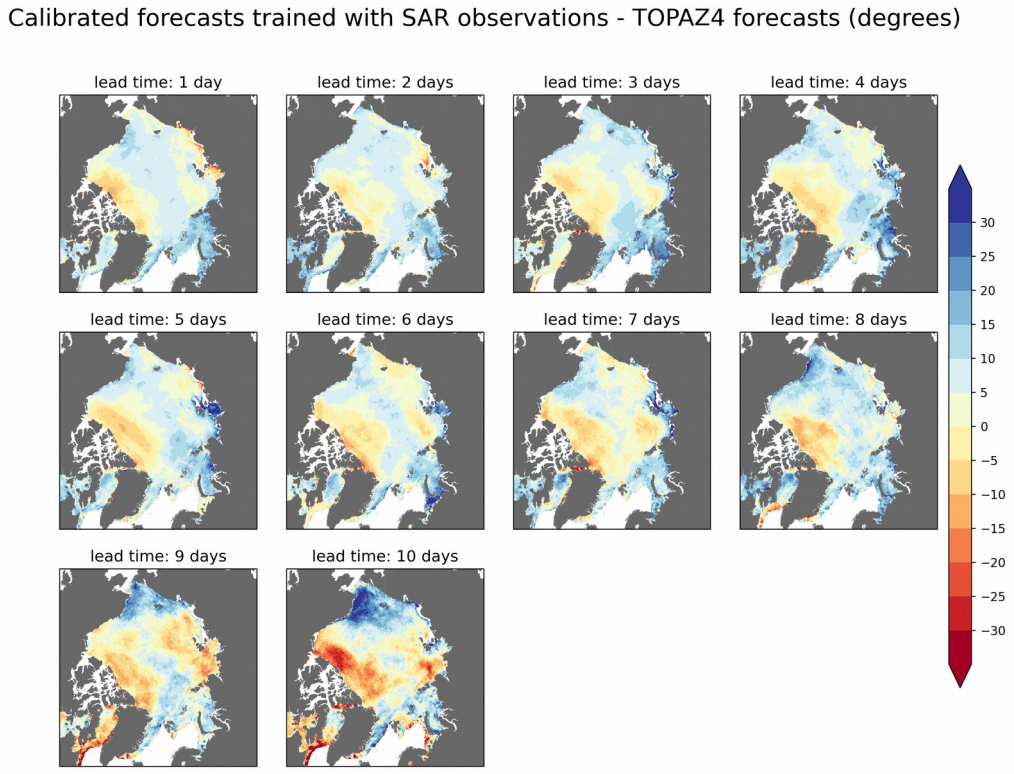


Figure S14. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

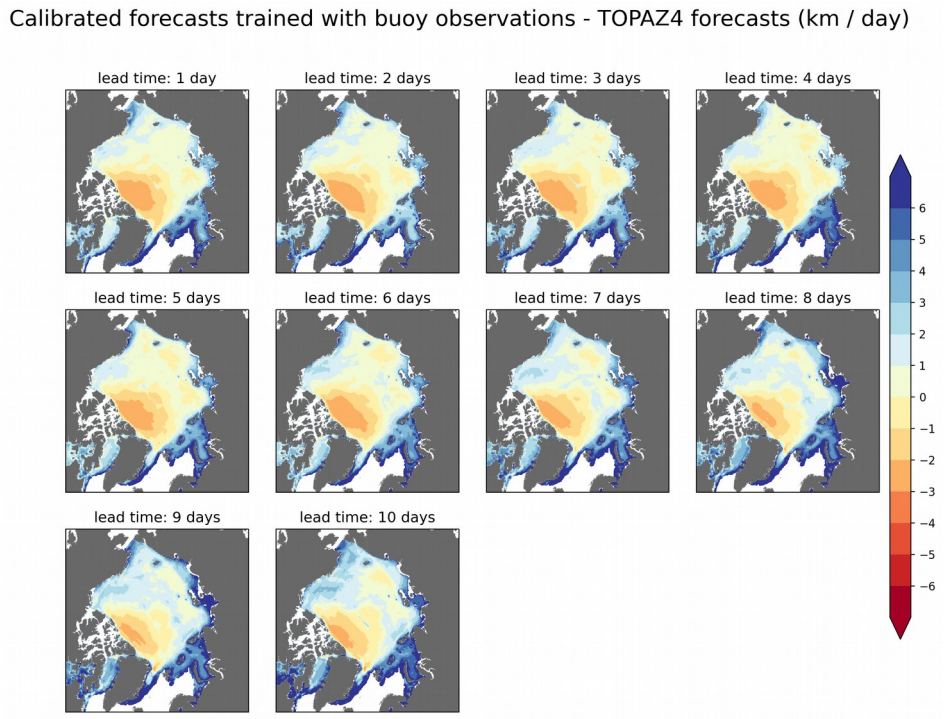


Figure S15. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

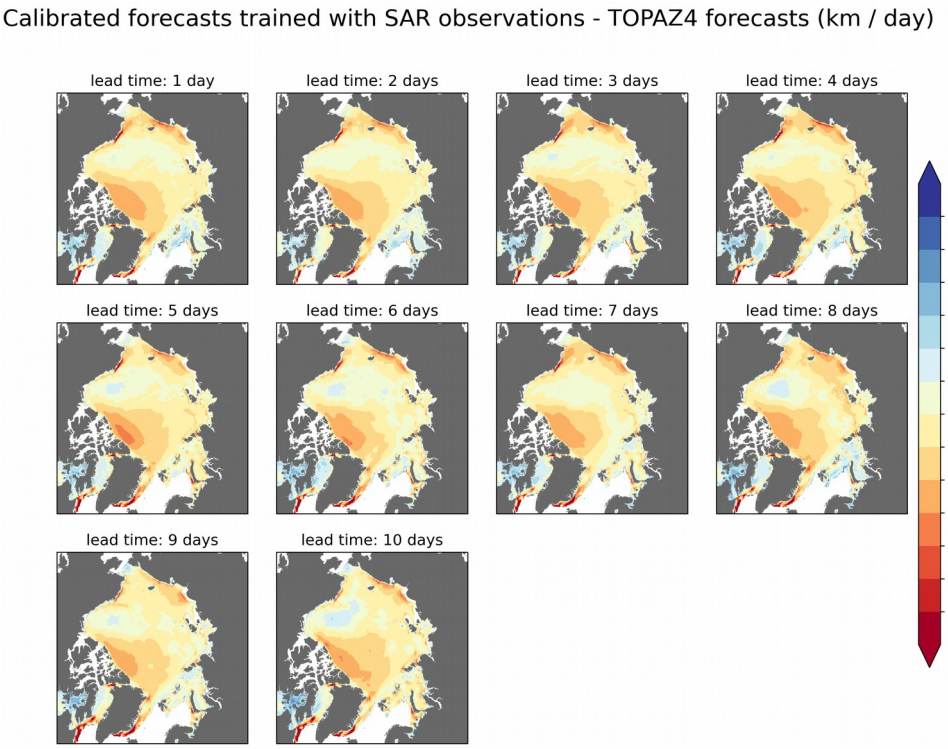


Figure S16. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

One general remark pertains to the Lagrangian nature of sea ice drift. The variable influencing the drift at a lead time of several days may not be at the same location as the sea ice drift value. This issue is not addressed in the paper, what do the authors expect to be the effect of considering both the predictor and the target at the same location?

We agree that the spatial variability in the sea-ice conditions plays a role in the sea-ice drift predictions. However, our approach consists of using the most reliable information available at the same location as the target variable. We think that this approach makes sense because the ECMWF and TOPAZ4 forecasts at a given location are influenced by the atmospheric and sea-ice conditions in the forecasts around this location. Furthermore, it would be possible to use a coarser resolution for the predictors. Nevertheless, this approach could be problematic in areas with a high spatial variability (for example near the coastlines). Another option would be to use several predictors for the same variable at different locations, but this would likely increase the risk of overfitting due to the spatial correlation between these predictor variables.

The authors have also neglected the seasonal changes of the forecast model performance, as well as the long-term model drift (or rather the absence of sea ice acceleration) as pointed out originally by Rampal et al. (2011) and then Xie et al. (2017) using an almost identical model.

We agree that using the period June-November 2020 for evaluating the forecasts was not ideal, and we have updated the results using the period from June 2020 to May 2021. We have also analyzed the sea ice drift trends in the buoy observations and in TOPAZ4 (figure S6 of the supplementary material):

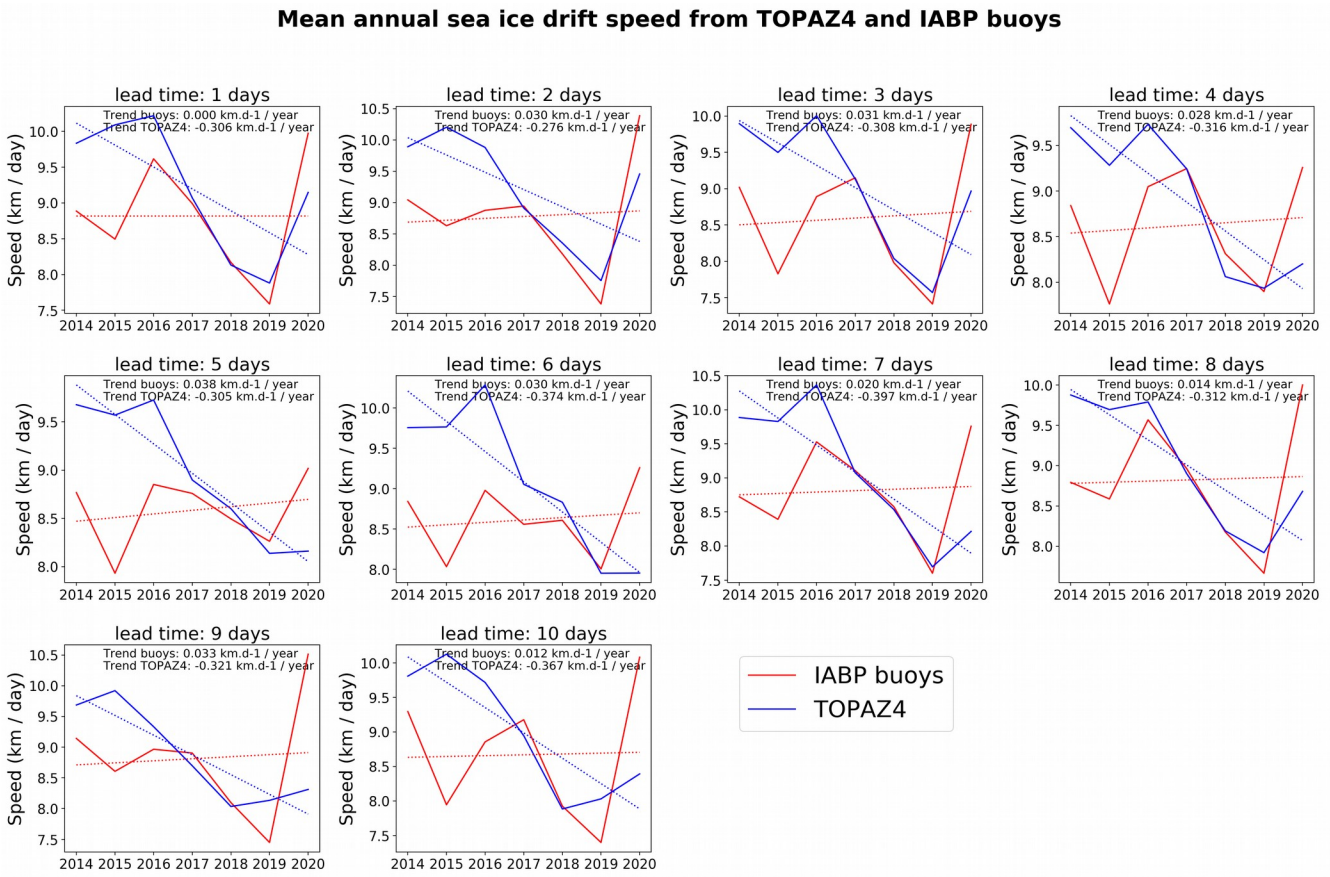


Figure S6. Mean annual sea ice drift speed from collocated IABP buoy observations and TOPAZ4 forecasts. The solid lines show the mean annual sea ice drift speed (km / day) from buoy observations and TOPAZ4 forecasts. The dashed lines show the linear trends.

We have added the following sentences in the section 3.2 of the revised version of the paper:

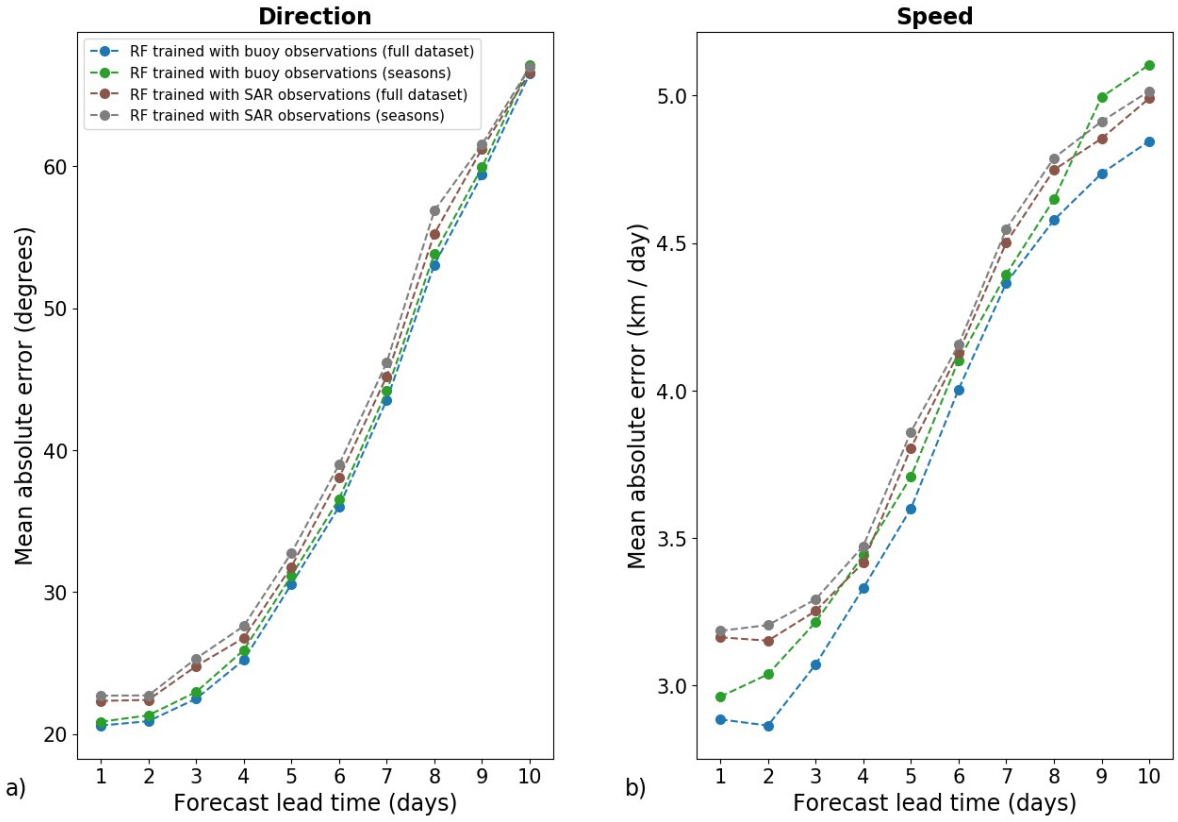
"Several training periods were tested between June 2012 and May 2020, and the chosen period from June 2013 to May 2020 seems to be optimal for predicting the direction of sea ice drift. However, using a shorter training period would have improved the forecasts for the speed of sea ice drift (figure S2 of the supplementary material). This is probably due to the smaller bias of TOPAZ4 sea ice drift speed in the recent years, which results from the negative trend of the sea-ice drift speed in TOPAZ4 (in contrast with IABP observations which show an acceleration, see figure S6 of the supplementary material)."

And the following sentence in the discussion and conclusion:

"Moreover, TOPAZ4 does not reproduce the recent acceleration of sea ice drift as already reported by Xie et al. (2017), and the bias of TOPAZ4 sea ice drift speed has changed during the studied period. This probably affects the performances of the random forest models trained with buoy observations due to their relatively long training period."

Can the algorithm learn the seasonality of the errors or could it be improved if trained separately on summer and winter data?

We have tested training the models separately on summer and winter data (see figure below), but this results in calibrated forecasts less accurate than developing only one model with the full training data set. Furthermore, we have also tested using the "day of year" as an additional predictor (see figure below). However, this results in a decrease in forecast accuracy, except for the random forest models predicting the speed of sea-ice drift which are trained using buoy observations. Based on these results, we have decided to discard the "day of year" from the list of predictors.



Comparison between the random forest (RF) models trained using the full training data set and the models trained separately for the winter and the summer. The mean absolute errors are assessed during the period from June 2020 to May 2021.

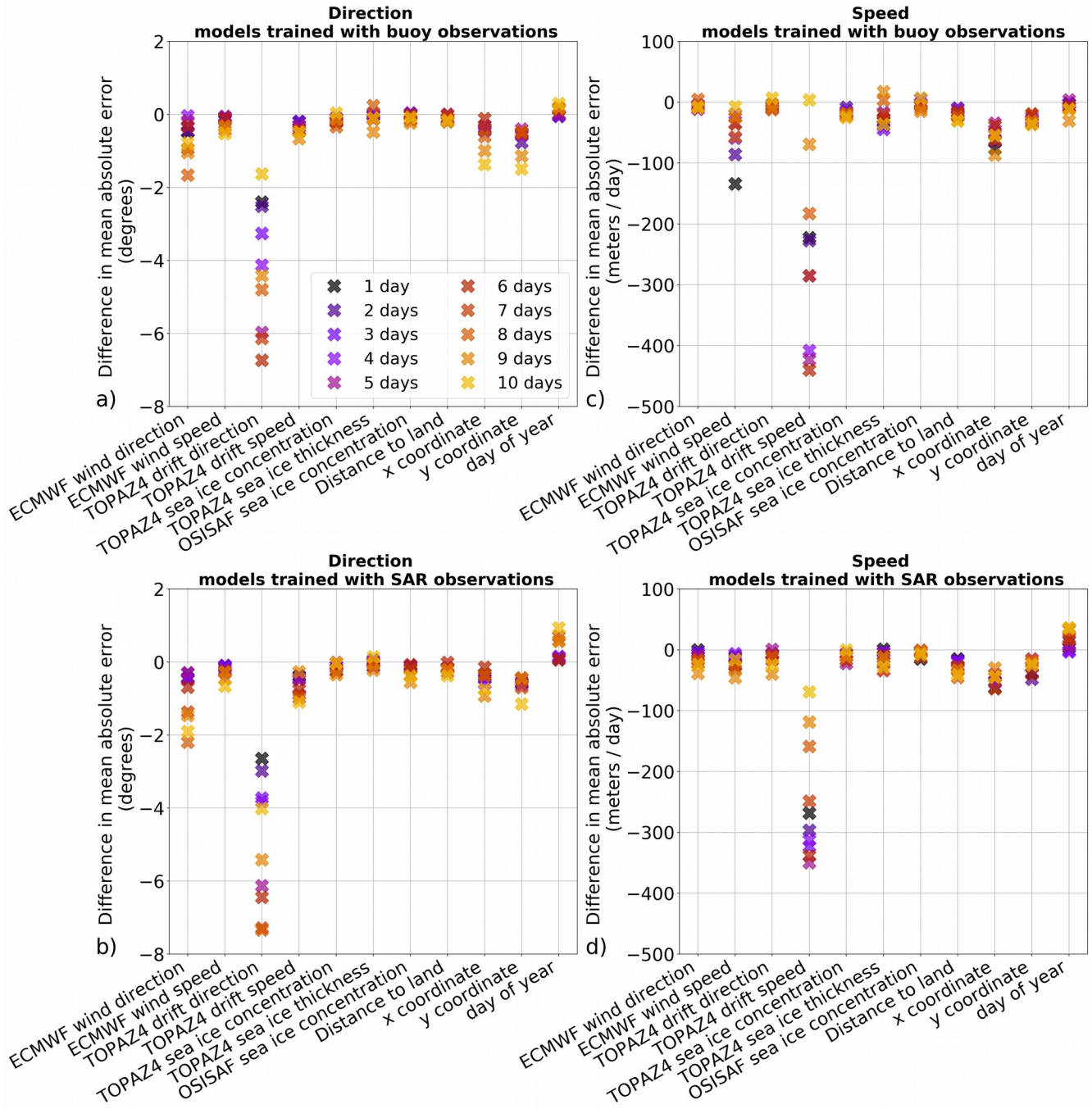


Figure S7. Differences in mean absolute error when one of the predictor variables is not used in the random forest algorithms for the direction (a, b) and speed (c, d) of sea-ice drift. The results are shown for the algorithms trained with buoy observations (a, c), and for the algorithms trained with SAR observations (b, d). The lead times are indicated in the legend of figure a). The differences represent the subtraction between the performances of the algorithms using all the predictor variables and the algorithms in which one predictor variable was not used. Therefore a negative value means that adding the variable in the algorithm improves the forecasts.

We have also added the following sentence in the section “4.3 Importance of predictor variables”:

"Furthermore, we also tested using the day of year as an additional predictor variable (figure S7 of the supplementary material), but adding this variable tends to deteriorate the forecast accuracy for most models, so we decided to discard this variable."

The manuscript cites the relevant literature and is original in its goals. I am not aware of any similar study carried out elsewhere. The article is logically structured and reads quite well. The figures are generally nice and clear. Exceptions are noted in detailed comments below.

Based on the above, I recommend the manuscript is published with minor corrections.

Detailed comments:

- **P1, l21: The relationship is complex and nonlinear in the ice pack where the rheology is active, but for low ice concentrations, the ice is in free drift and should be a linear function of the winds (the Nansen relationship).**

Thanks for this comment. We have added the following statement in the introduction: *"Though sea ice drift is mainly driven by the wind in areas with a low sea ice concentration, the relationships between these variables and sea ice drift are complex and not linear in most of the ice-covered areas (Yu et al., 2020)."*

- **P2, l29: "but they obtained": false opposition. Is there any reason why RF or CNNs would have an advantage for sea ice concentrations?**

The authors suggest that it might be due to the larger learning capacity of the CNN model compared to the RF model, in particular concerning the ability of CNN to learn spatial features from the predictors. The following sentence:

"Recently, Kim et al. (2020) developed sea-ice concentration forecasts based on random forests and convolutional neural networks, but they obtained more accurate results using convolutional neural networks."

has been replaced by:

"Recently, Kim et al. (2020) developed and compared sea-ice concentration forecasts based on random forests and convolutional neural networks. They obtained more accurate results using convolutional neural networks probably due to the larger learning capacity of convolutional neural networks compared to random forests, in particular to extract spatial features from the predictors (Kim et al., 2020)."

- **P3, l78: The overestimation of sea ice drift was reported in reanalysis, but since the decadal acceleration of sea ice drift is not reproduced by the model, the bias should be smaller in recent times, as can be seen in the TOPAZ4 validation pages: https://cmems.met.no/ARC-MFC/V2Validation/timeSeriesResults/year-day-01/SItimeSeries_year-day-01.html#drift (accessed 2nd March 2021)**

We have analyzed the sea ice drift trends in the buoy observations and in TOPAZ4 (figure S6 of the supplementary material):

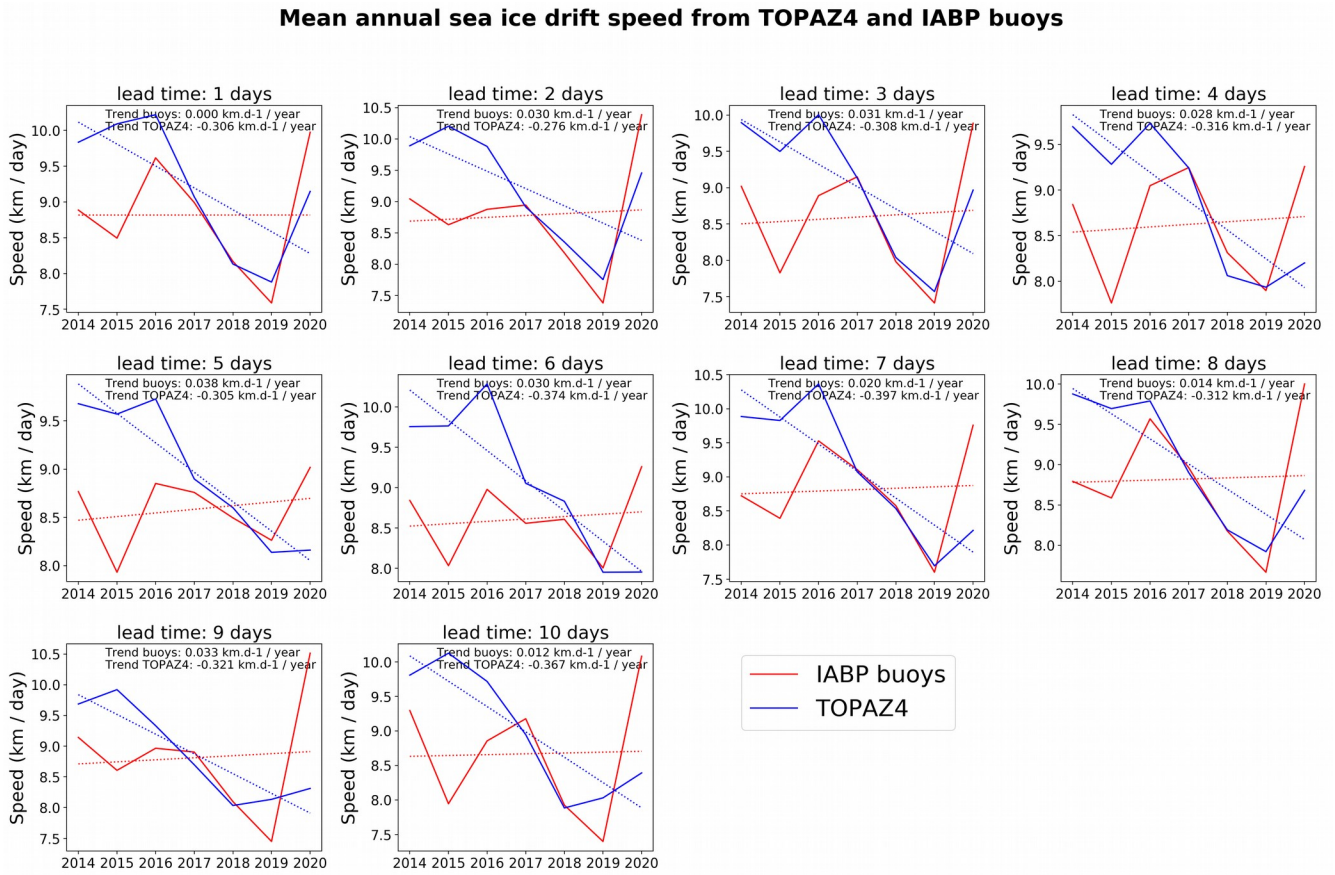


Figure S6. Mean annual sea ice drift speed from collocated IABP buoy observations and TOPAZ4 forecasts. The solid lines show the mean annual sea ice drift speed (km / day) from buoy observations and TOPAZ4 forecasts. The dashed lines show the linear trends.

We have added the following sentences in the section 3.2 of the revised version of the paper:

"Several training periods were tested between June 2012 and May 2020, and the chosen period from June 2013 to May 2020 seems to be optimal for predicting the direction of sea ice drift. However, using a shorter training period would have improved the forecasts for the speed of sea ice drift (figure S2 of the supplementary material). This is probably due to the smaller bias of TOPAZ4 sea ice drift speed in the recent years, which results from the negative trend of the sea-ice drift speed in TOPAZ4 (in contrast with IABP observations which show an acceleration, see figure S6 of the supplementary material)."

And the following sentence in the discussion and conclusion:

"Moreover, TOPAZ4 does not reproduce the recent acceleration of sea ice drift as already reported by Xie et al. (2017), and the bias of TOPAZ4 sea ice drift speed has changed during the studied period (figure S6 of the supplementary material). This probably affects the performances of the random forest models trained with buoy observations due to their relatively long training period."

- **l108: "different algorithms were used": "models" should not be synonymous with "algorithm" (the Random Forest is one algorithm, from which you can build several models). Maybe use "distincts models were developed to..."?**

We agree with this comment and we have replaced "algorithms" by "models" here and several times in the paper.

- **L148: At which point is the averaging used? Is it related to the averaging of each prediction tree?**
  The prediction from a random forest model used for regression is the mean value of the predictions from all decision trees. In our study, the decision trees predict an angle in degrees (between 0 and 360 °). These angles are then converted to complex numbers in order to average the angles predicted by the decision trees. The mean value from all decision trees (the final prediction) is then converted in degrees. In order to clarify this point, we have replaced *"results"* by *"predictions" (line 147 of the discussion paper),* and we have added *"(in degrees)"* in the following sentence:
  "In order to avoid this issue, the predictions from all decision trees (in degrees) were converted to complex numbers before averaging."

The new paragraph:
"The prediction from a random forest model used for regression is the mean value of the predictions from all decision trees. For the direction of sea ice drift, each decision tree predicts a value between 0 and 360°. When averaging several predictions close to the northward direction, this can be an issue because values slightly higher than 0° and slightly lower than 360° can be averaged, possibly leading to a mean value close to the southward direction. In order to avoid this issue, the predictions from all decision trees (in degrees) were converted to complex numbers before averaging. Then, the average of complex numbers was converted into an angle in degrees."

- **L148 If the predictive variable is a complex number, isn't it similar to predict normalised u and v components (with a norm of 1)? In that case, this choice is apparently contradictory with the assertion line 90: "In order to predict independent variables, it has been chosen to forecast the direction and speed of sea-ice drift rather than the eastward and northward components"**

Because the complex numbers are only used to average the angles predicted by all decision trees (in degrees), we do not think that this is similar to predict the normalized u and v components.

- **Section 3.2: It is very positive that sensitivity studies are detailed. The algorithms were tuned against the size of the training set (period for buoys, subsampling rate for SAR), size of the forest (number of trees), other parameters of the RF. It is not clear to me which criteria were used for this tuning. On which dataset the error has been computed to evaluate the tuning? Is it the one used to evaluate the results (buoys in June-November 2020) or the one used to evaluate the importance of predictor variables (section 4.3)?**

We have added a supplementary material in which the results from the sensitivity experiments are described. We have decided to use only the training period for all the sensitivity experiments (similar as section 4.3), except for the period used for training the random forest models using buoy observations because this does not make sense for this one. We have added the following paragraph in the method section (section 3.1):

"In order to optimize some parameters of the algorithms, sensitivity tests were performed using only data from the training periods (see supplementary material). For these sensitivity tests, the random forest models were trained using data from about 80 % of the forecast start dates (randomly selected) within the training periods. Then, the data from the remaining forecast start dates were used for evaluating the forecast performances. This selection prevents using neighboring grid points with very similar conditions in the training and validation data sets, and was repeated 10 times in order to obtain robust results. Furthermore, the random forest models were evaluated using the same product as the one used for training for these sensitivity tests (CMEMS SAR MOSAIC

product for those trained with SAR observations, and IABP buoys for those trained with buoy observations). This method was also used to evaluate the optimal fraction of the grid points covered by SAR observations used for training some random forest models (see section 3.2), as well as to assess the importance of the predictor variables (see section 3.5).”

- **L183: The period chosen for evaluating the model is mostly in the summer season (June-November)? Do you expect it to be representative of the winter? The link above shows a seasonal signal in the drift bias, though not a large one.**

We agree that using the period June-November 2020 for evaluating the forecasts was not ideal, and we have updated the results using the period from June 2020 to May 2021.

- **L206-2018. It is fair to note the absence of data where the performance deteriorates. This however deserves an explanation as to how the random forest algorithm extrapolates the training data spatially. Does it find the most analogous situations where and when training observations are available? The authors explain that the random forest does provide the average of an ensemble but it would be good to have insights about the values returned, for example, in places of intermittent landfast ice.**

We hope that the section 3.1 of the revised manuscript will help to understand this better. In this section, the principle of random forest algorithms is described (see below). Basically, the decision trees will find the most analogous situation depending on the predictor variables chosen to split the nodes. Furthermore, we have removed the Canadian Archipelago from our analysis in order to reduce the issues related to the presence of landfast ice.

“Random forest algorithms consist of an ensemble of decision trees used for regression or classification tasks (Breiman, 2001). In order to avoid overfitting (meaning that the models learn from noise in the training data), independent decision trees must be developed. The independence of decision trees is ensured by using different subsets of the training data set for developing each decision tree, as well as by randomly selecting a fraction of the predictor variables at each node (the node is then split using the variable maximizing a dissimilarity metric among the selected predictors). Each decision tree is trained with a data set created using the bootstrap method, which consists of randomly selecting samples from the original training data with replacement for creating a new data set of the same size as the original one. This results in using about 63 % of the samples from the original data set for training each decision tree.
In this study, random forest models were developed for regression using the Python library Scikit-learn-0.23.2 (Pedregosa et al., 2011), and the mean squared error was used to measure the quality of the splits.”

- **L224-226 "The selection of the data sets used for training and evaluating the random forest models is a random process according to the forecast start date to avoid the influence of neighboring grid points with very similar conditions," this point of correlations between training and validation data (leading to data leakage) is essential to avoid correlation between training/validation data that could lead to data leakage and overfitting. It would be beneficial for the community to give more details (even if it is given in appendix) about your selection procedure.**

We have added the following paragraph in section 3.1:
“*In order to optimize some parameters of the algorithms, sensitivity tests were performed using only data from the training periods (see supplementary material). For these sensitivity tests, the random forest models were*

trained using data from about 80 % of the forecast start dates (randomly selected) within the training periods. Then, the data from the remaining forecast start dates were used for evaluating the forecast performances. This selection prevents using neighboring grid points with very similar conditions in the training and validation data sets, and was repeated 10 times in order to obtain robust results. Furthermore, the random forest models were evaluated using the same product as the one used for training for these sensitivity tests (CMEMS SAR MOSAIC product for those trained with SAR observations, and IABP buoys for those trained with buoy observations). This method was also used to evaluate the optimal fraction of the grid points covered by SAR observations used for training some random forest models (see section 3.2), as well as to assess the importance of the predictor variables (see section 3.5)."

- **l 236: Intuitively one may expect that the areas of thicker ice drift slower than thin ice due to the increased resistance to stress.**

We have added the following statement in the section "4.3 Importance of predictor variables":

"Furthermore, the mean absolute errors for the speed of sea ice drift are also considerably reduced by adding the sea ice thickness forecasts from TOPAZ4 (between 0.011 and 0.098 km / day), probably due to the anti-correlation between sea ice thickness and sea ice drift speed (Yu et al., 2020)."

- **Section 4.3. This sensitivity study is important. But I am surprised not to see the standard "Importance variable" diagnostic available in any random Forest algorithm? Even if the results are redundant with your study, it would have offered another point of view of variable importance.**

We have added an analysis of the relative importance of the predictors using the impurity-based feature importance method (figure 8 of the revised version of the paper):

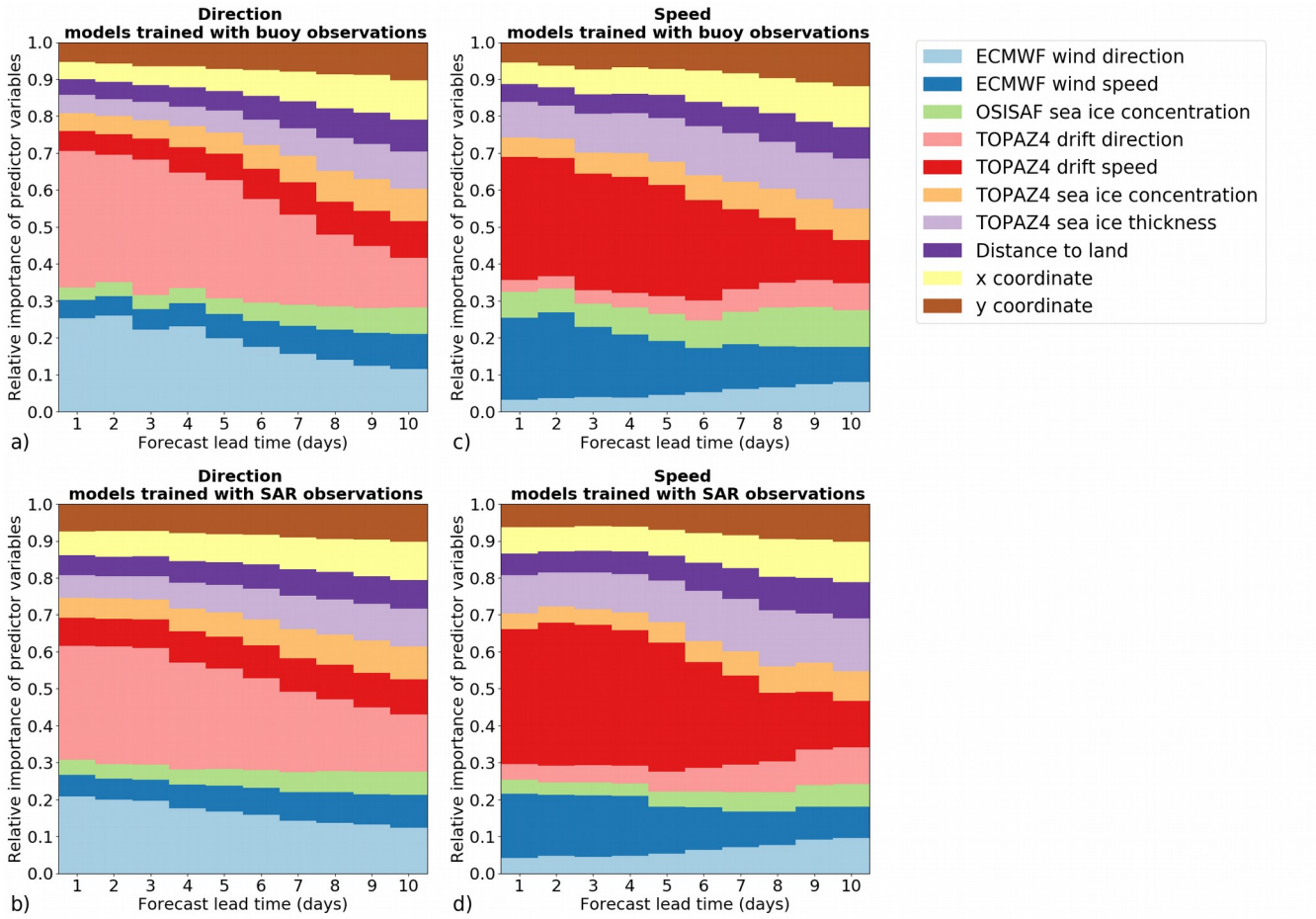


Figure 8. Relative importance of the predictor variables for the direction (a, b) and the speed (c, d) of sea ice drift assessed using the impurity-based feature importance method.

We have added the following paragraph in the section "3.5 Evaluation of the importance of predictor variables": *"In this study, the importance of the predictor variables was estimated using two different methods. First, the impurity-based feature importance was assessed. This method is based on the measure of impurity decreases (the mean squared error here) at all nodes in the random forest algorithm (the variables that often split nodes with large impurity decreases are considered important). It provides an assessment of the relative importance of the predictor variables, but is known for underestimating the importance of non-continuous predictors (Strobl et al., 2007)."*

And the following paragraphs in the section "4.3 Importance of predictor variables":

"For both calibration methods, the most important variable for predicting the drift direction is the sea ice drift direction fromTOPAZ4 forecasts, followed by the wind direction from ECMWF forecasts (figure 8). On average, the relative importance of sea ice drift direction forecasts is about 1.4 and 1.5 times larger than the one from wind direction forecasts for the models trained with buoy and SAR observations, respectively (figure 8). The sum of the relative importances of these two variables represent, on average, about 46 and 41 % of the sum of all relative importances for the models trained with buoy and SAR observations, respectively. However, the relative importances of these two variables decrease with increasing lead times.

Similarly, the sea ice drift speed from TOPAZ4 is the most important variable for predicting the speed of sea ice drift, followed by the wind speed from ECMWF forecasts. On average, the relative importance of sea-ice drift speed forecasts is about 1.7 and 2.2 larger than the one from wind speed forecasts for the models trained with buoy and SAR observations, respectively (figure 8). For the models predicting the speed of sea ice drift, the sum of the relative importances of these two variables represent, on average, about 40 % of the sum of all relative importances for both calibration methods. Furthermore, the relative importances of these two variables also decrease with increasing lead times."

- **L258: It is correct to mention the changes in operational systems but the authors should note that even with unchanged reanalysis systems, the gradual acceleration of ice drift is not reproduced by the models and may also affect the training over long periods.**

  We have added the following statement in the discussion and conclusion section:
  "*Moreover, TOPAZ4 does not reproduce the recent acceleration of sea ice drift as already reported by Xie et al. (2017), and the bias of TOPAZ4 sea ice drift speed has changed during the studied period (figure S6 of the supplementary material). This probably affects the performances of the random forest models trained with buoy observations due to their relatively long training period***"**

- **L261: I may have misunderstood this point. I do not expect any 7-days frequency signal in sea ice drift so Thursdays are representative of the rest of the week.**

Our point here is that TOPAZ4 forecasts could be more accurate when they start on Thursdays than on other days due to data assimilation. If so, it means that the weights given to the different predictors might not be optimal for the forecasts not starting on Thursdays. The ECMWF wind forecasts and the sea-ice concentration observations could have larger weights if daily forecasts would have been used for training the random forest algorithms. In order to clarify this, the following sentence:

"Because only the forecasts starting on Thursdays are initialized using data assimilation, this could be an issue when producing forecasts not starting on Thursdays."

has been replaced by:

"Because data assimilation is only performed on Thursdays, this could be an issue when producing forecasts not starting on Thursdays (the weights of the different predictor variables might not be optimal)."

- **Code availability: I would like to point out that there is not enough details given on the results so it can be reproduced. It is said that "the codes used for this analysis can be made available upon request." but without the code, it is not possible to reproduce the results as the RF models are not detailed.**

We have created a github directory (https://github.com/cyrilpalerme/Calibration_of_sea_ice_drift_forecasts/) in which the codes are available.

- **Figures 2 and 10: the crosses colours are not colourblind-friendly. Try a simpler scale - a gradient - that can easily distinguish the high from the low percentages. The general tendency is more interesting to me than the exact values.**

Thanks for this comment, we have changed the color scale of these figures.

- **Figures 4 and 5: do we need to see both the MAE and the RMSE ?**

We agree with this comment and we have removed the RMSE.

References:

Rampal, P., Weiss, J., Dubois, C. and Campin, J.-M.: IPCC climate models do not capture Arctic sea ice drift acceleration: Consequences in terms of projected sea ice thinning and decline, J. Geophys. Res., 116, C00D07, doi:10.1029/2011JC007110, 2011.

Xie, J., Bertino, L., Counillon, F., Lisæter, K. A. and Sakov, P.: Quality assessment of the TOPAZ4 reanalysis in the Arctic over the period 1991–2013, Ocean Sci., 13(1), 123–144, doi:10.5194/os-13-123-2017, 2017.