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 1

##########

 # Summary

Palerme and Müller use random forest regression to predict Arctic sea-ice drift speed and direction from a set of predictors that contains besides dynamical sea-ice drift forecasts (TOPAZ4) also wind forecasts, geographical coordinates, sea-ice concentration and thickness, and distance from land. Using both buoy and satellite-derived drift for training and evaluation, the authors find that the predicted drift slightly outperforms the original TOPAZ4 drift forecasts at all lead times considered (1-10 days); mean absolute errors are reduced by roughly 5-10%. In my view the study is very relevant and innovative, scientifically sound, and well presented. What I think deserves additional effort is to illuminate more clearly what happens within the "black box" of the random forecast algorithm, for example, which of the predictands are picked how often to split nodes, what the output resolution of the individual trees is, how the predictands "modify" the TOPAZ4 drift forecasts, how that compares to simpler bias corrections, and how such characteristics change with lead time. With more explanations along these lines, the article could help readers (including myself) to better understand how the approach really functions, thereby providing an educational example how ML methods can help us to enhance predictions beyond the direct outputs of numerical models. In summary, I recommend publication of this work in The Cryosphere subject to minor(-to-major) revisions as detailed in the following.

##########

 # Specific comments

Regarding the term "calibration": In my view it would be helpful to clarify in how far the presented approach is a "calibration" of dynamical model-based drift forecasts. Typically, calibration in this context means to use raw dynamical model forecasts and to modify them in some systematic way, e.g., to remove model biases. However, here the TOPAZ4 drift forecasts are used qualitatively in the same way as the other predictands, which appears to be a conceptual deviation from the standard calibration approach and leads to interesting questions. For example, would there be ways to formulate the random forecast algorithms such that they are explicitly used to modify the raw TOPAZ4 drift forecasts rather than predicting the drift "from scratch"? Or is that basically equivalent to the way it's currently being done, treating the TOPAZ4 drift just like any other predictand? It would be good to provide some clarification and/or discussion in this regard.

All the predictors are similarly provided to the random forest algorithms (it is not possible to explicitly define some predictors as more important than others before training the models). However, the most relevant predictors will be used more often to split the nodes, and will have a more important role in the predictions than other predictors. Furthermore, we agree that the meaning of the term "calibration" here differs from systematic bias corrections. Nevertheless, it is commonly used to describe weather forecasts produced using machine learning techniques, including random forests based on similar approaches as our study (for example: Gagne et al., 2014 ; Loken et al., 2019 ; Hill et al., 2020). Therefore, we have decided to keep the term "calibration" in the manuscript.

P2L47+56: "... have been used for training some random forest algorithms ...": First, from these sentences it is at first not clear that you are not talking about previous work, but that this is what has been done in the present study. Second, the "some" sounds very vague, maybe you can refer here to Sect. 3.2.

We agree that it was not clear in the text, and we have replaced these sentences by:
"In this study, satellite sea ice drift observations from the CMEMS product named SEAICE_GLO_SEAICE_L4_NRT_OBSERVATIONS_011_00675 (MOSAIC version 2.0, hereafter referred as CMEMS SAR MOSAIC product) were used for training some random forest algorithms (see section 3), as well as for analyzing the spatial variability of the performances of sea ice drift forecasts."

and

"In addition, data from the International Arctic Buoy Programme (IABP) were also used for training some random forest algorithms (see section 3), as well as for evaluating the SAR observations and the sea ice drift forecasts."

Sect. 2.2.: I think it would help to make very clear here that the TOPAZ4 drift forecasts are the basic ingredient here, but that other predictands are added and actually treated in the same way as the TOPAZ4 drift forecasts within the random forest algorithms, see my previous remarks.

A more detailed description of the random forest method has been added in section 3.1 of the revised version of the paper which describes how the predictor variables are selected to split the nodes:

"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 (Pedregosaet al., 2011), and the mean squared error was used to measure the quality of the splits. Different models were developed for predicting the direction and speed of sea ice drift, as well as for different lead times (1 to 10 days). Moreover, two sets of models were developed using target variables either from buoy displacements or from SAR observations. Therefore, 20 different models were developed using buoy displacements, and 20 other models were developed using SAR observations. "

P3L79-80: "while TOPAZ4 forecasts are produced daily, only the forecasts starting on Thursdays are initialized using data assimilation": This sounds as if the forecasts starting on other days than Thursdays would not at all be affected by data assimilation, but I assume that they are affected by previous data assimilation, that is, from the last Thursday (and earlier), right? So I would say they are still "initialized", just not with particularly timely observations.

It is right that the TOPAZ4 forecasts are affected by the previous data assimilation (last Thursday). The sentence:

"However, while TOPAZ4 forecasts are produced daily, only the forecasts starting on Thursdays are initialized using data assimilation and stored in the long-term archive."

has been replaced by:

“While TOPAZ4 forecasts are produced daily, data assimilation is only performed on Thursdays, and only the forecasts starting on Thursdays are stored in the long-term archive.”

P4L91: "The initial bearing on the great-circle path": From the context one can guess what is meant by "bearing" here, but is this word really correct?

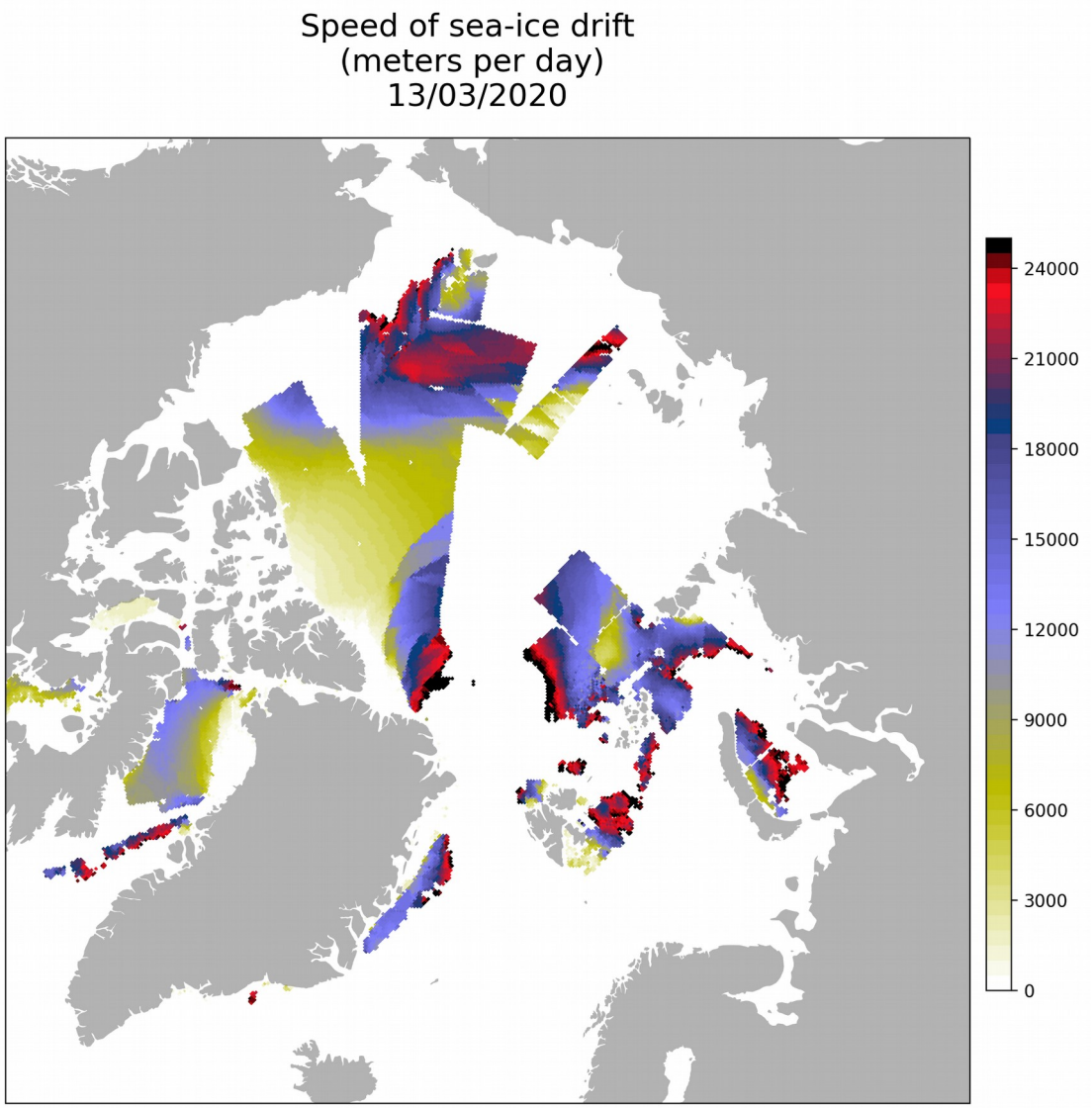
Thanks for this comment. We have checked this, and *“initial great-circle course angle”* seems to be the most common term. We have used this term in the revised version of the paper.

P5L120: "as independent data sets": Please clarify what you mean here exactly by "independent".

We meant that we used all the grid points with buoy observations similarly for training the random forest algorithms. However, some of the grid points could be spatially correlated, and the term “independent” would not be appropriate. We have decided to remove the term “independent” here.

P5L121-133: Given that, if I understand correctly, the main motivation for subsetting the SAR data is to avoid the use of highly-correlated neighbouring data points and thus overfitting, wouldn't it be more effective to do the thinning in a more systematic way by omitting more points in data-rich regions rather than subselecting completely randomly without taking data density into account?

The spatial distribution of the number of SAR observations is influenced by the orbit of the satellites and by the sea-ice extent. Therefore, the spatial distribution of the number of observations used for training the algorithms shown in figure 1 c) is influenced by the seasonal cycle of the sea-ice extent. Furthermore, there is a high variability in the spatial coverage of the MOSAICs (see example below), and some regions can be well covered during a particular day while there are not many observations in these regions during the full training period. Nevertheless, the grid points in these regions can be highly correlated, and a sub-sampling can be necessary. The regions with many observations (typically the Central Arctic) are also the regions with the most reliable observations due to a larger number of overpasses. Therefore, reducing the number of grid points used in the Central Arctic could potentially reduce the quality of the observations used as target variables, and having a negative impact on the random forest algorithms. Though we consider this question as very interesting and relevant, we also think that this is a complex question which is out of the scope of our paper (which is a first attempt of using random forests for calibrating sea-ice drift forecasts). Therefore, we have decided to keep the method which consists of randomly selecting the grid point covered by SAR observations.



Example of MOSAIC showing the speed of sea-ice drift on 13/03/2020 from the CMEMS product named SEAICE_GLO_SEAICE_L4_NRT_OBSERVATIONS_011_006 (MOSAIC version 2.0).

P5L130: By evaluating only over the period June-November 2020, doesn't this potentially introduce a seasonal bias for the evaluation? (This also raises the question whether it would be worthwile considering to add the time of the year as an additional predictand?)

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. Furthermore, we have 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. We have added the figure below in the supplementary material and the following sentence in the main paper (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"

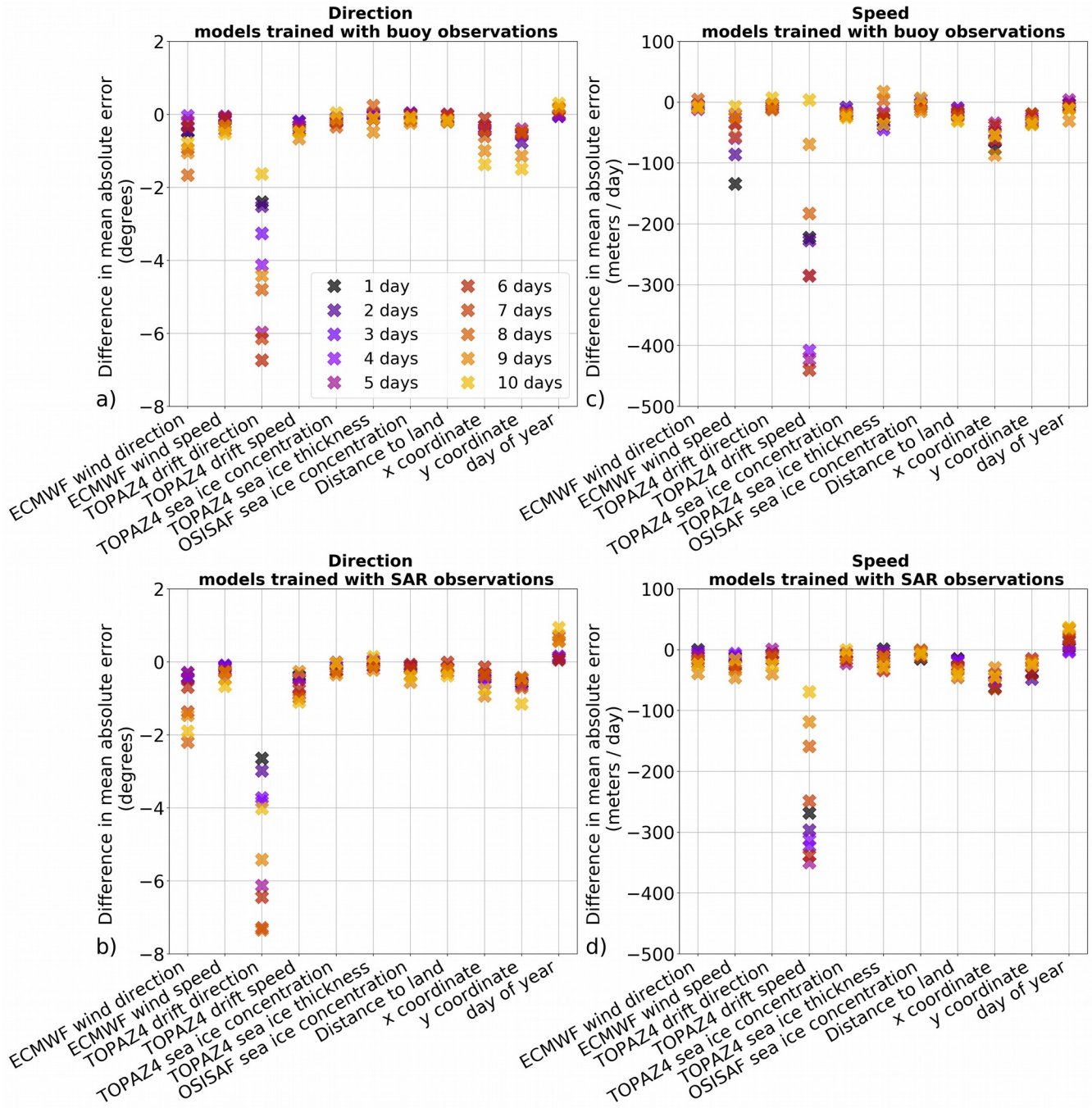


Figure S7. Differences in mean absolute error when one of the predictor variables is not used in the random forest algorithms for the direction (a, b) and speed (c, d) of sea-ice drift. The results are shown for the algorithms trained with buoy observations (a, c), and for the algorithms trained with 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.

P5L133: "10^4 training data sets": Should this be "data points"?

We agree that “data sets” can be confusing, and we have replaced it by “data points”.

P5L143: Here again, the TOPAZ4 drift forecasts are mentioned just alongside all other predictands - shouldn't they he highlighted much more upfront as the "main predictors" (which are to be "calibrated")?

We think that TOPAZ4 drift forecasts should not be highlighted as the main predictors in this section because all the predictor variables are provided similarly to the random forest algorithms. Furthermore, we have added a paragraph in section 3.1 of the revised paper to better describe the random forest method:

“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. Different models were developed for predicting the direction and speed of sea ice drift, as well as for different lead times (1 to 10 days). Moreover, two sets of models were developed using target variables either from buoy displacements or from SAR observations. Therefore, 20 different models were developed using buoy displacements, and 20 other models were developed using SAR observations. 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 algorithms were evaluated using the same product as the one used for training the random forest models 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). Based on the sensitivity tests, we decided to develop random forest models using 200 decision trees (there were no significant improvements when using more trees), to maximize the depth of the decision trees (most of the leaves contain only one sample from the training data set), and to set the number of predictor variables considered for splitting the nodes at three. These parameters were chosen for all the models developed”

P5L143: Also, I think it would be good to state clearly that for a specific lead time only the forecasts (TOPAZ4 & IFS) for that specific lead time are used as predictands - or is that not the case?

This is true that it was not mentioned in the text. We have modified the following sentence:
“The variables from sea-ice and wind forecasts during the predicted lead time can be considered as the last category.”

P6L153-155: "maximizing the depth of the decision trees" - First, given that the decision trees are based on quasi continuous predictor variables as well as continuous target variables, there does not appear to be an absolute "maximum" depth. Can you please specify what depth is actually used? Second, related to this, how meany leaves do the individual decision trees have, and how are the associated predicted values distributed?

We hope that the section 3.1 of the revised manuscript will help to understand this better (see our previous responses). The leaves of a decision tree must contain at least one sample from the training data set obtained after bootstrapping. By maximizing the depth of the decision trees, we develop decision trees in which most of the leaves contain only one sample from the training data set. Due to bootstrapping, the number of leaves is about 63 % of the size of the original training data set. However, it can happen that the target variable has the same value multiple times, and that the associated predictors are very similar (for example with very correlated grid points). This explains why some leaves can have several samples, even when maximizing the depth of the decision trees. Therefore the number of leaves is not fixed, and can vary slightly (though close to 63 % of the size of the original training data set in our study).

Furthermore, the depth of a decision tree is not fixed in our study and varies depending on various parameters such as the size of the training data set (which varies depending on lead time, and the bootstrap method also slightly influences the number of independent samples), as well as the structure of the tree (not all leaves are at the same depth).

Do the resulting distribution densities approximately match the distributions of the target variables (or does the "resolution" vary in a specific way)?

The distribution densities of the decision trees match the distribution of the target variable during the training period. However, the predicted value from a random forest model is the average of the predictions from all decision trees, which tends to reduce the number of extreme values predicted. In our study, this should not be an issue for predicting the direction (due to the circular nature of directional data), but this could be an issue for the speed of sea ice drift.

We have added the following sentences in the section “3.1 Development of random forest models”:

“Furthermore, random forest models tend to predict less extreme values than the target variable because the mean value from all decision trees is used as the prediction. This should not be an issue for predicting the direction of sea ice drift due to the circular nature of directional data, but particularly low and high sea ice drift speed could be difficult to predict with random forest models.”

P6L153-155: "setting the number of predictor variables considered for splitting the nodes at three": First, I speculate this small number of random predictands per split "forces" the algorithms to use the less-informative predictands (other than TOPAZ4 drift and IFS winds) more often than a decision tree would do that can always choose from all predictands. Can you provide some more insight into this?

We have added the following paragraph in section 3.1:
“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."

Second, related, which predictands are chosen how often to split nodes? I imagine over a large number of layers, TOPAZ4 drift (or IFS winds) would always be preferred over other predictands as long as those main predictands are not yet used so often that the resulting resolution of the target variable is approximately as high as the effective accuracy of those forecasts in the first place. Do you find such a systematic behaviour, that the "main" predictands dominate the upper layers and "other" predictands gain importance in lower layers?

The random selection of the predictors at each node makes this analysis biased. Even predictors that are not very important are sometimes chosen to split a node in the upper layers of a decision tree because they are the most effective predictor among the selected predictors. Instead of analyzing how often the predictors are chosen to split the nodes, we have decided to add an analysis of the impurity-based feature importance. This method is more commonly used than analyzing the number of times each predictor is selected by all individual trees in the forest, and considered more robust (e.g. Strobl et al., 2007). This analysis is shown in a new figure (figure 8 in the revised paper, see response to the next comment), and we have added the following paragraph to explain this method:

"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)."

Moreover, does the relative "use frequency" of different predictands change for the different lead times? For example, I could imagine that the relative importance of TOPAZ4 drift versus winds might change with lead time, which might in turn be related to the way IFS forcing and perturbations are used to drive the ice and ocean in TOPAZ4?

We have analyzed the evolution of the relative importances of the predictors over lead times using the impurity-based feature importance method (figure 8 of the revised version of the paper):

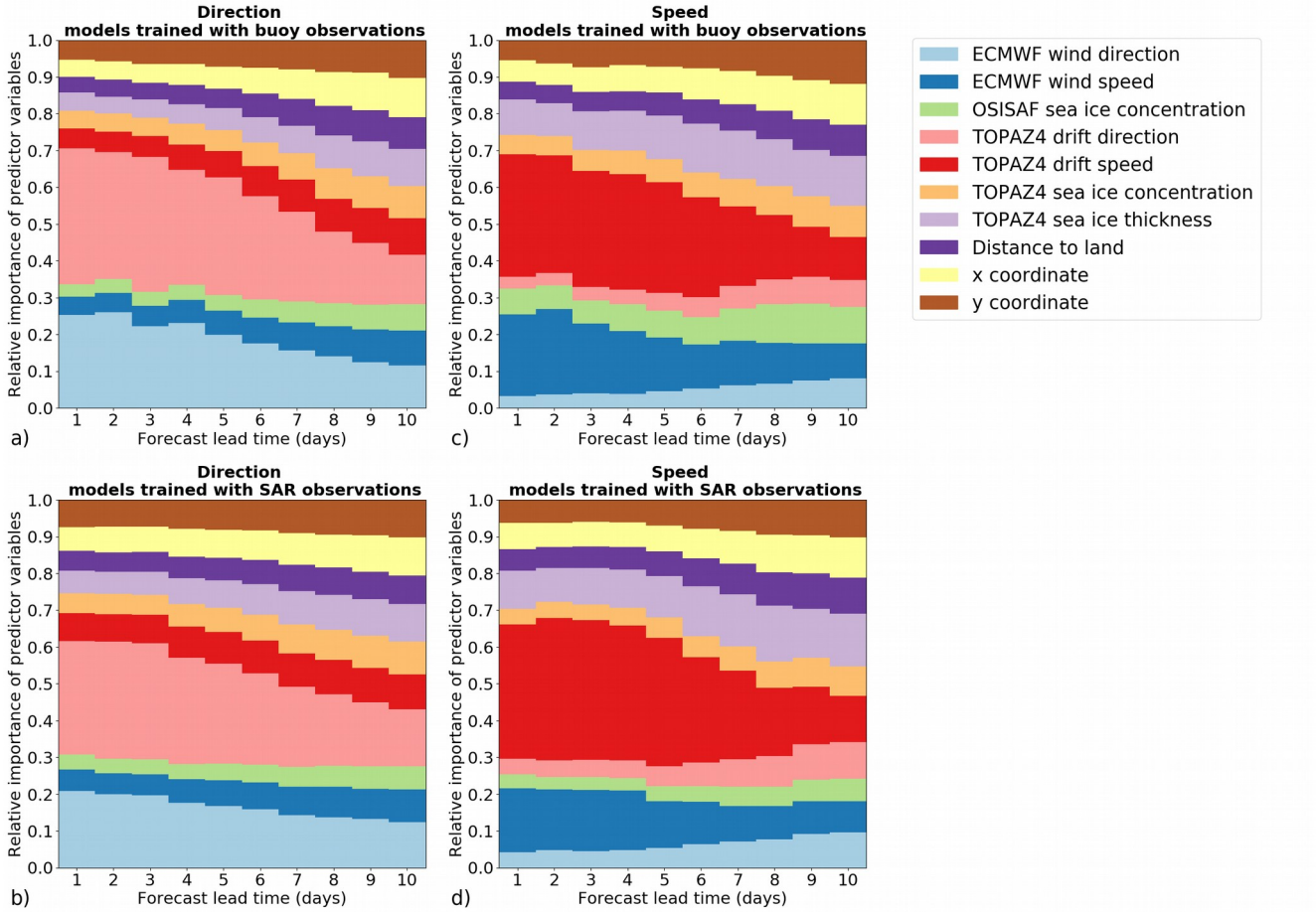


Figure 8. Relative importance of the predictor variables for the direction (a, b) and the speed (c, d) of sea ice drift assessed using the impurity-based feature importance method.

Sect. 4.3: First of all, I really like these sensitivity experiments to quantify the impacts of individual predictors. As mentioned above, I think it would be really helpful to add more information about how often the predictands are actually used in the regression trees, which I suppose would provide similar information about relative importance from a very different angle - in fact without the need to run additional algorithms.

We have answered to this comment in our previous two responses.

Furthermore, it is not surprising that the TOPAZ4 drift forecasts (speed for speed, direction for direction) are the most important predictands, right? Again, this makes me wonder how the approach followed here relates to classical "calibration", that is, to use a raw forecast and "modify" it based on some additional information, and how the final forecasts derived here deviate from the raw forecasts. E.g., are the raw drift speeds and directions systematically corrected (on average) in one or the other way - and maybe this depends on the region (e.g., CAA vs. open ocean), the lead time, and the sea-ice thickness or concentration? Some more information and discussion regarding these aspects would in my view be very helpful.

We have added the figures below which show the mean differences between the calibrated forecasts and TOPAZ4 forecasts in the supplementary material. We have answered to the rest of this question in our response to the next comment.

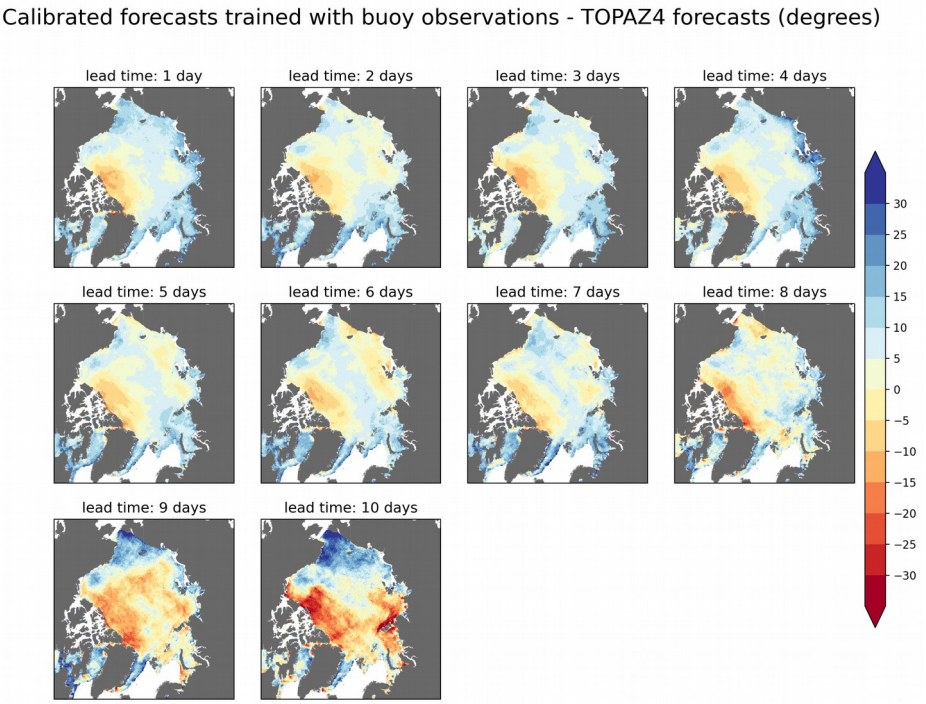


Figure S13. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

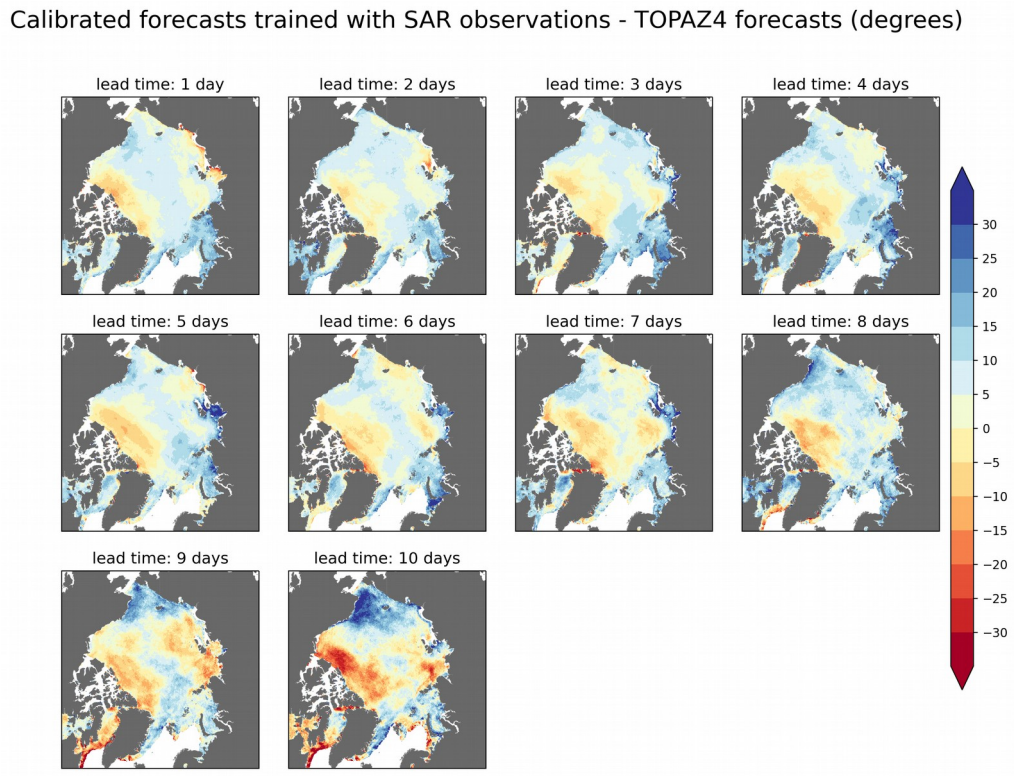


Figure S14. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

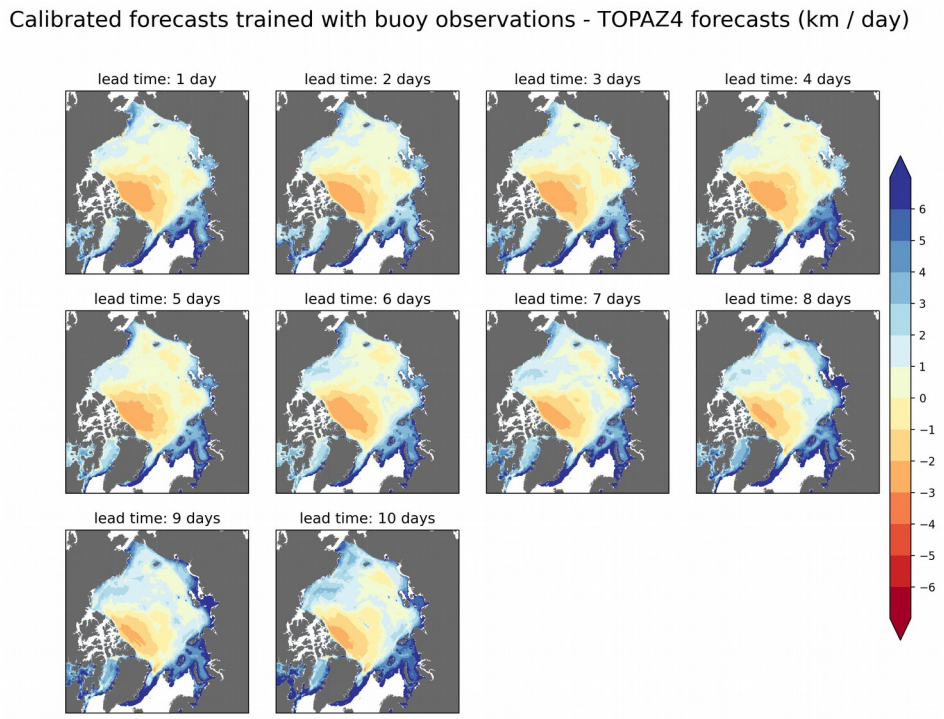


Figure S15. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

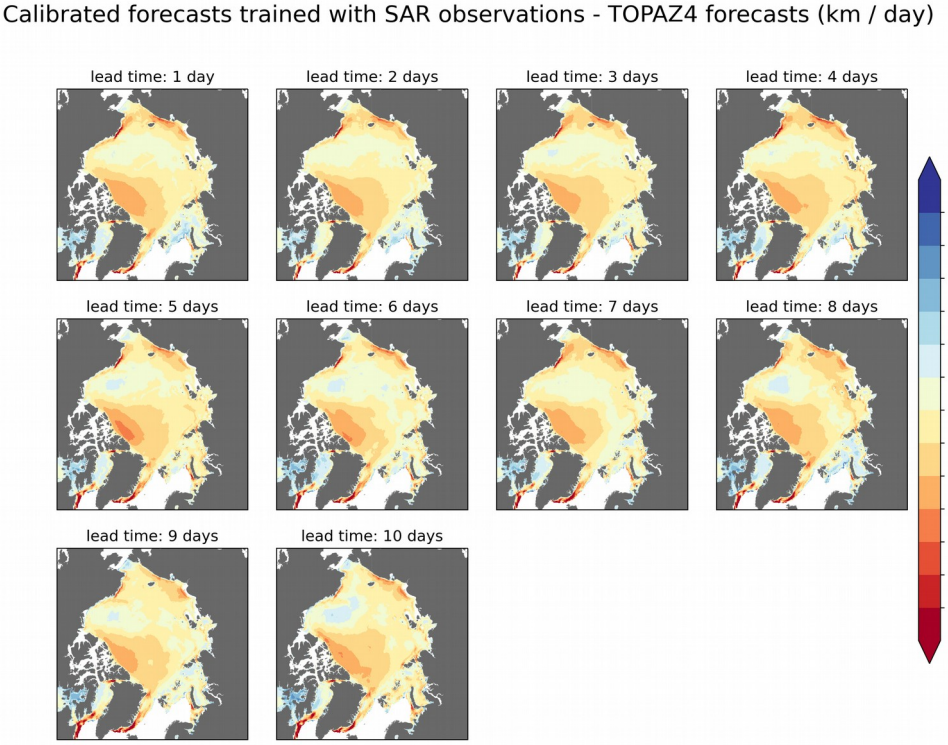


Figure S16. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

Following up on the previous point(s), I am wondering in how far similar improvements (over raw TOPAZ4 drift) might have been achieved with a simpler ("classical") calibration approach, e.g., by correcting the drift speeds and directions with some constant factors and/or offsets? In this regard, it would also he helpful to see if mean biases for speed and direction exist that could be corrected for by such a trivial calibration approach. On the other hand, if such simple biases are absent, that might be a strong argument against such simplistic calibration, right?

We have added some maps of the TOPAZ4 biases during the period 2018 – 2019 compared to SAR observations in the supplementary material:

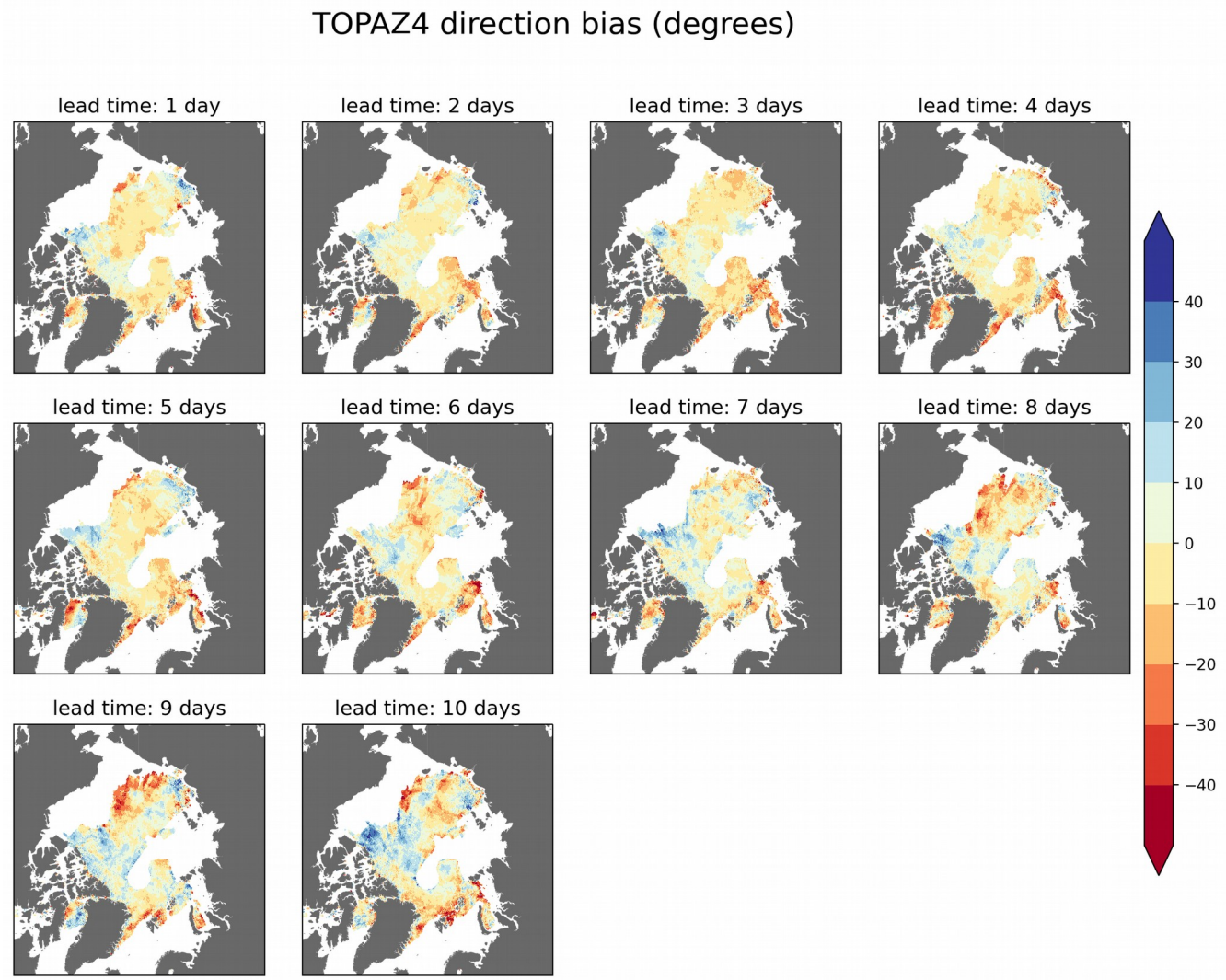


Figure S9. Mean direction bias (degrees) from TOPAZ4 forecasts during the period 2018 - 2019 compared to SAR observations. Only the grid points containing at least 20 SAR observations during the period 2018 - 2019 have been taken into account.

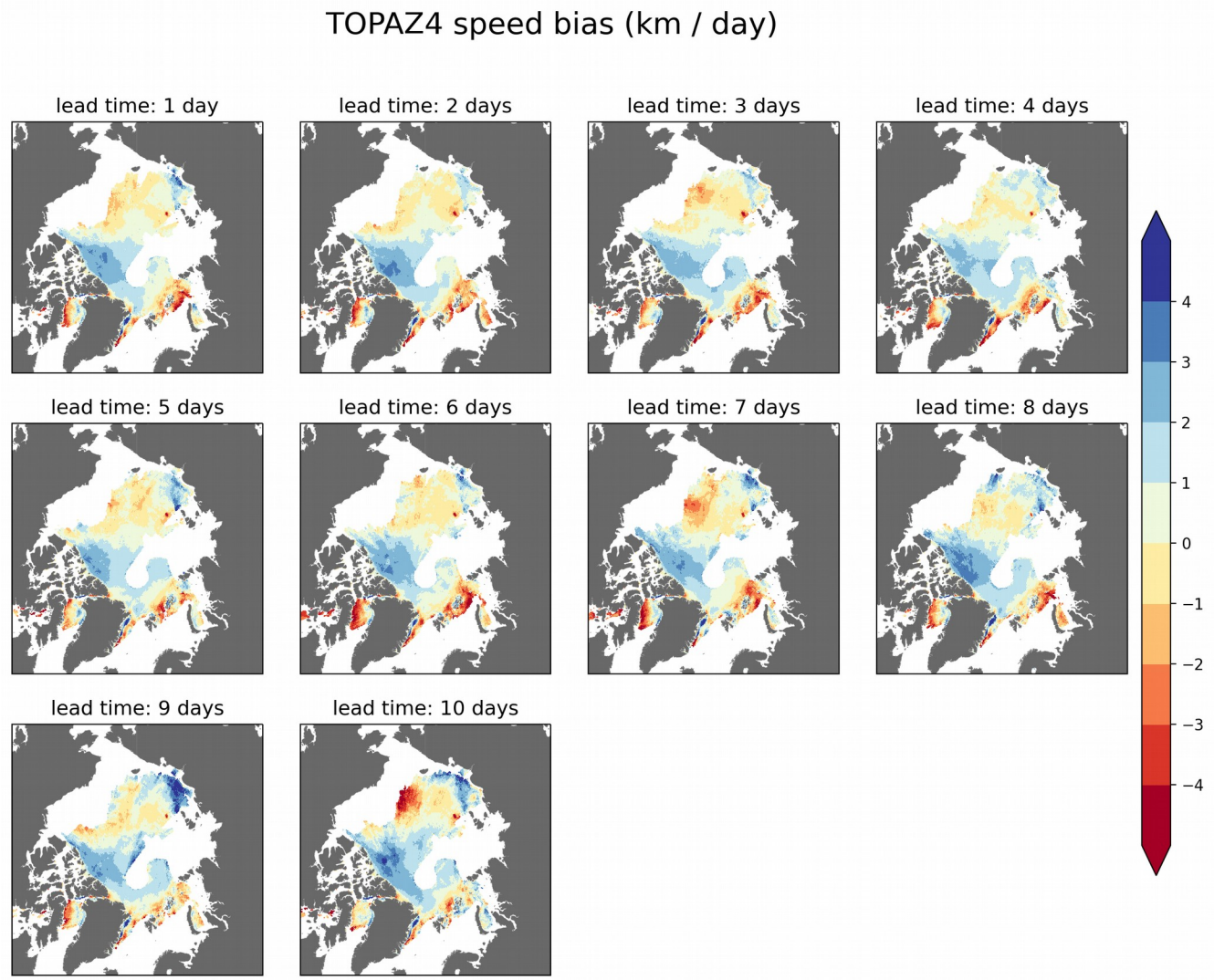


Figure S10. Mean speed bias (km / day) from TOPAZ4 forecasts during the period 2018 - 2019 compared to SAR observations. Only the grid points containing at least 20 SAR observations during the period 2018 - 2019 have been taken into account.

We have also compared the random forest models described in the paper with calibrated forecasts produced using a simple bias correction, as well as with random forest models using only 3 predictors: the spatial coordinates (x and y) and the predicted variable from TOPAZ4 (TOPAZ4 drift direction and speed for the models predicting the direction and speed of sea ice drift, respectively). For the bias correction of TOPAZ4, we have calculated the bias during the period 2018-2019 for all grid points with at least 20 SAR observations for a given lead time. Therefore, the biases are calculated using SAR observations as reference. Note that the limited number of available sea ice drift observations makes this approach limited because it is not possible to cover the entire Arctic (see figures above). Furthermore, due to this limited coverage, a smaller number of buoys have been used for evaluating the mean absolute errors of the forecasts in the figure below (figure S12 of the supplementary material) than in figure 3 of the main paper. We have also added the following paragraphs in the supplementary material:

2 Bias correction of TOPAZ4 forecasts

In order to compare the random forest models developed in this study with more simple calibration methods, we have developed calibrated forecasts by correcting the biases from TOPAZ4 forecasts. The biases from TOPAZ4 forecasts have been evaluated for each grid point and each lead time during the period 2018 - 2019 using SAR observations as reference (figures S9, S10, and S11). Only the grid points containing at least 20 SAR observations during the period 2018 - 2019 have been used for this calibration and for the evaluation presented in figure S12. For the direction of sea ice drift, the bias from the period 2018 – 2019 has been subtracted from the TOPAZ4 forecasts. For the speed of sea ice drift, the TOPAZ4 forecasts have been multiplied by the ratio of the SAR observations over the TOPAZ4 forecasts during the period 2018 – 2019.

3 Random forest models using only three predictors (figure S12)

Random forest models using only three predictor variables have been developed and compared to the other calibration methods. Only the x and y coordinates, as well as the drift direction from TOPAZ4 have been used for the models predicting the direction of sea ice drift. For the models predicting the speed of sea ice drift, the drift speed from TOPAZ4 has been used with the x and y coordinates. Note that the number of predictors randomly selected at each node has been fixed at two for these random forest models.

4 Random forest models predicting sea ice drift along the x and y axes of TOPAZ4 grid

We developed random forest models predicting the sea ice drift along the x and y axes of the TOPAZ4 grid using a different set of predictor variables (figure S12). For these models, the northward and eastward components of the ECMWF wind forecasts were used as predictors instead of the wind speed and direction, as well as the sea ice drift along the x and y axes from TOPAZ4 forecasts (which are provided by TOPAZ4 outputs) instead of the sea ice drift speed and direction. The direction and speed of sea ice drift were then calculated using the start and end location of the sea ice for comparing those models with the ones directly predicting the direction and speed of sea ice drift.

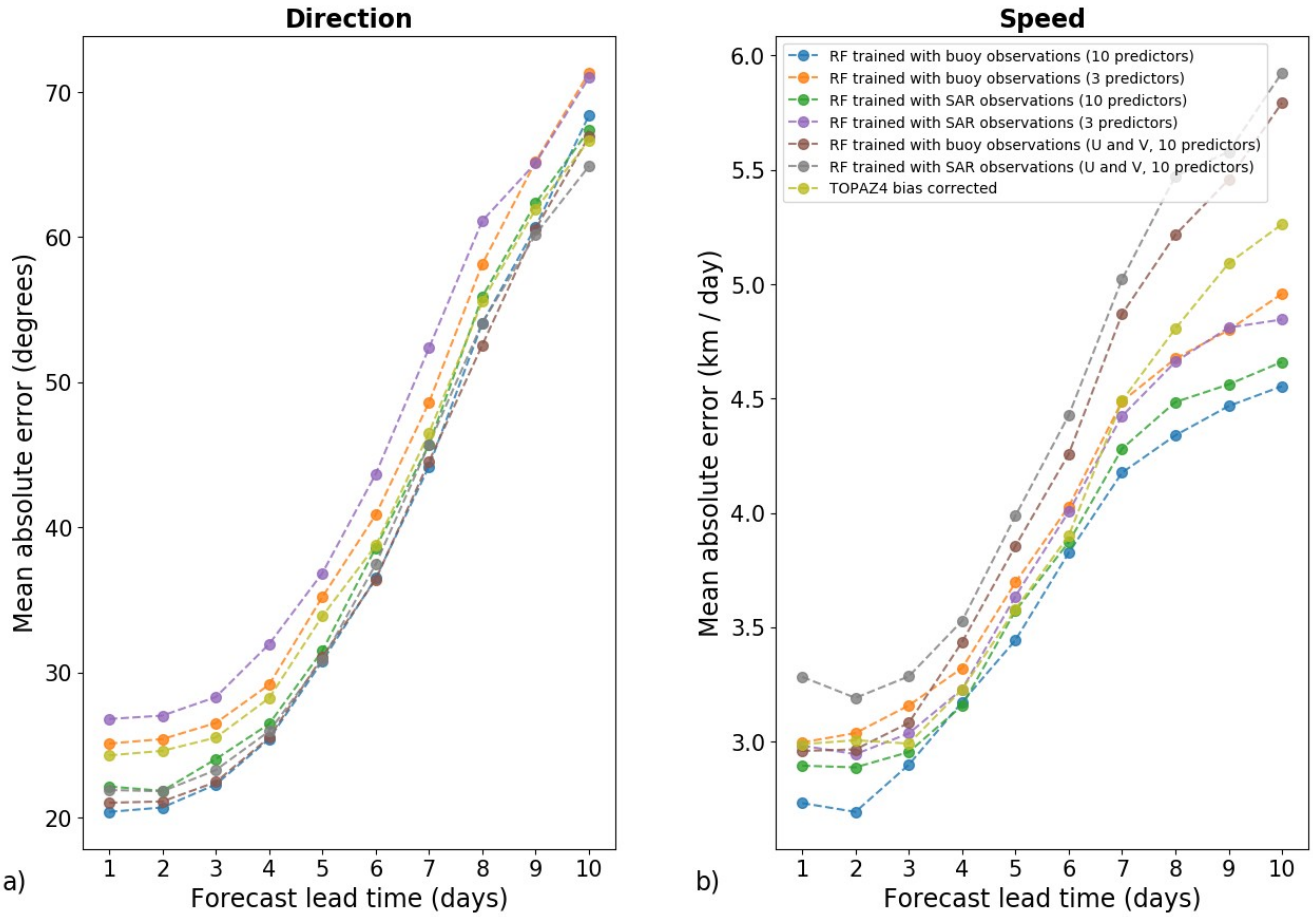


Figure S12. Mean absolute errors of different calibration methods for the period June 2020 – May 2021. Buoy observations have been used as reference. The random forest (RF) models using all the predictors (10 predictors) described in the main paper are shown by the blue and green curves. The random forest models using only 3 predictors (x and y coordinates, as well as the drift direction from TOPAZ4 for the models predicting the direction, and the drift speed from TOPAZ4 for the models predicting the speed of sea ice drift) are shown by the orange and purple curves. The TOPAZ4 forecasts which are bias corrected (using the period 2018-2019 for calculating TOPAZ4 bias) are shown by the yellow curve. The random forest models predicting the sea ice drift along the x and y axes of TOPAZ4 grid are shown by the brown and gray curves.

References:

Gagne, D. J., McGovern, A., and Xue, M.: Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts, Weather and Forecasting, 29, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1, 2014.

Hill, A. J., Herman, G. R., & Schumacher, R. S. (2020). Forecasting Severe Weather with Random Forests, *Monthly Weather Review*, *148*(5), 2135-2161.

Loken, E. D., Clark, A. J., McGovern, A., Flora, M., and Knopfmeier, K.: Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests, Weather and Forecasting, 34, 2017–2044, https://doi.org/10.1175/WAF-D-19-0109.1, 2019.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC bioinformatics, 8, 1–21, 2007.

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:

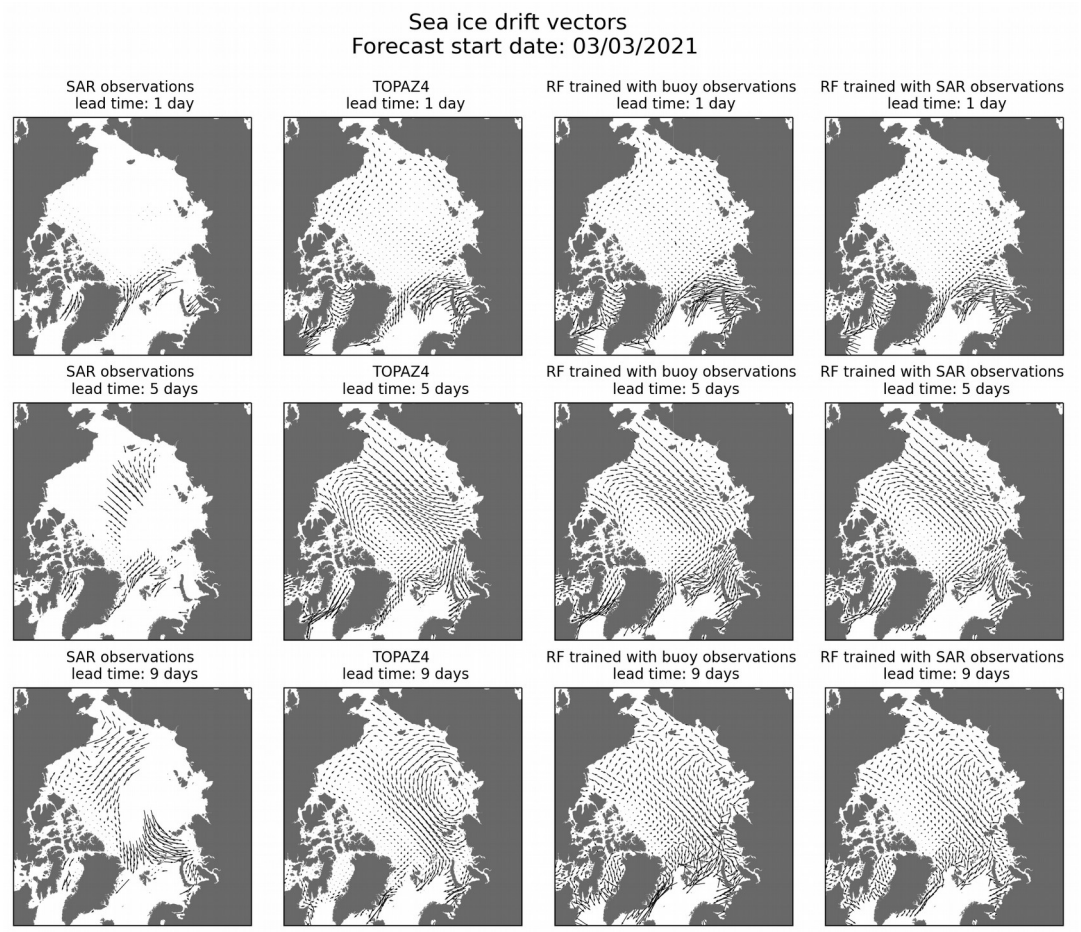


Figure S8. Example of calibration with the random forest (RF) algorithms for the forecasts which started on 03/03/2021 and for lead times of 1, 5, and 9 days.

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

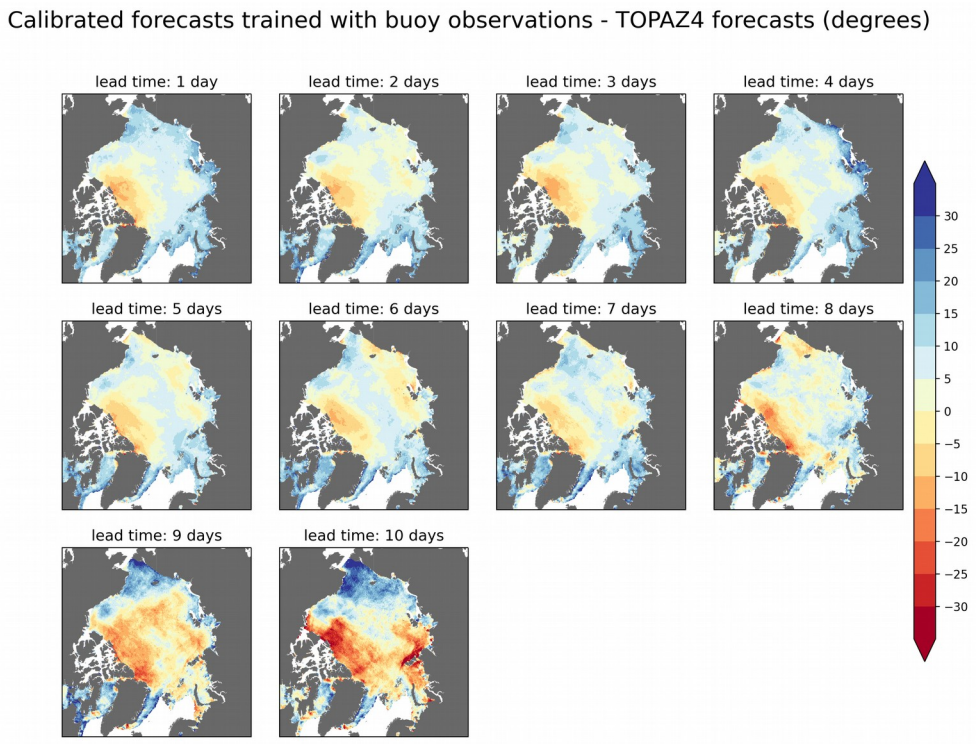


Figure S13. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

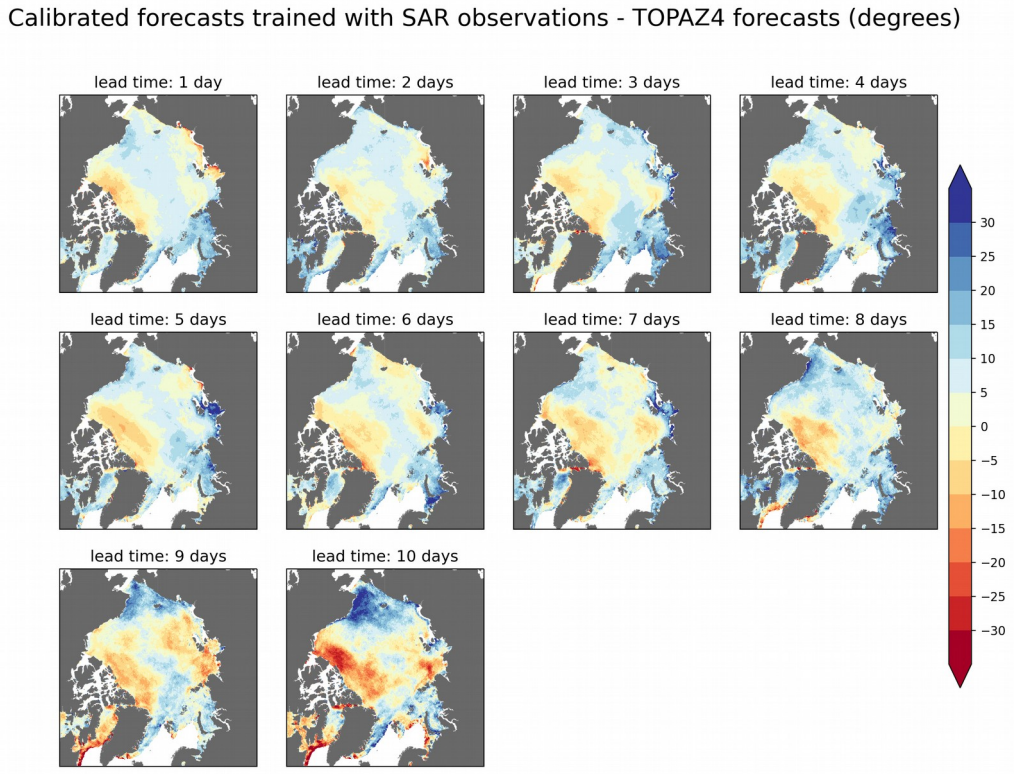


Figure S14. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

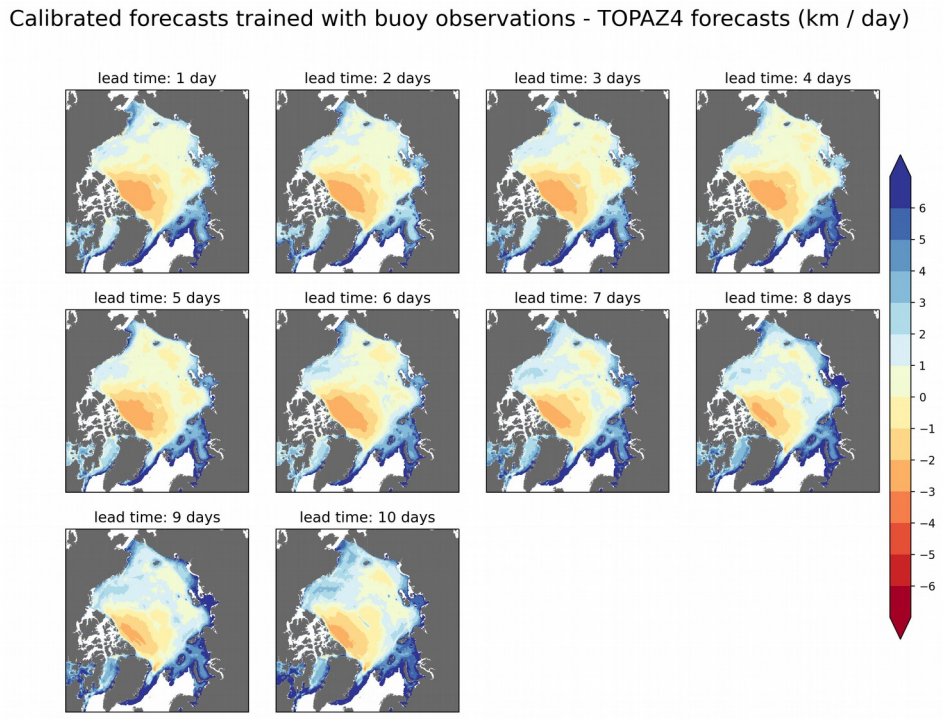


Figure S15. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

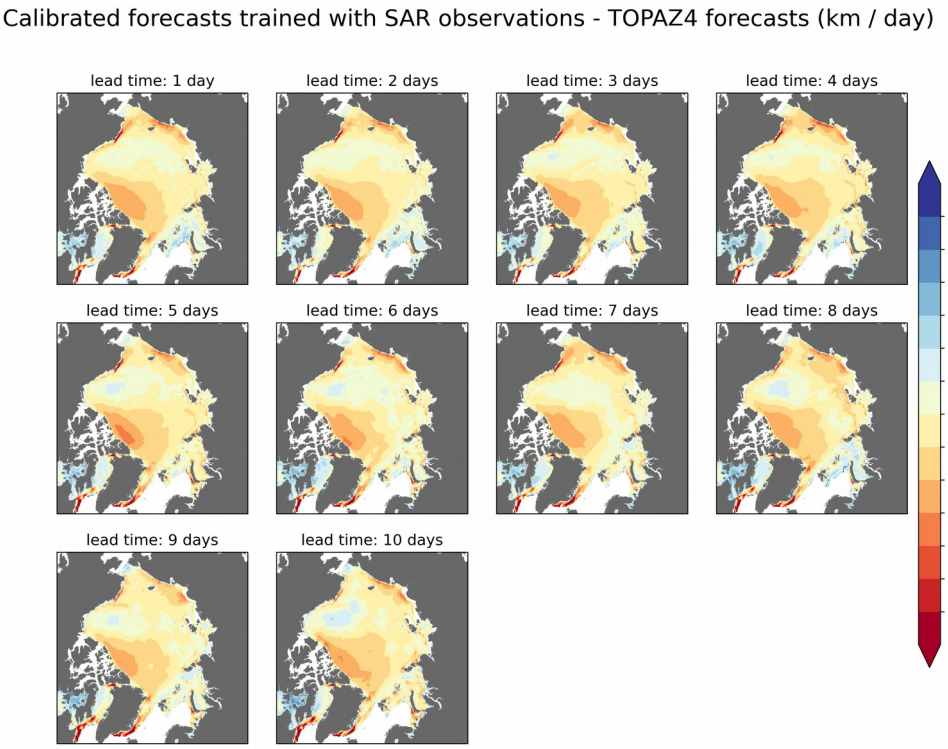


Figure S16. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

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):

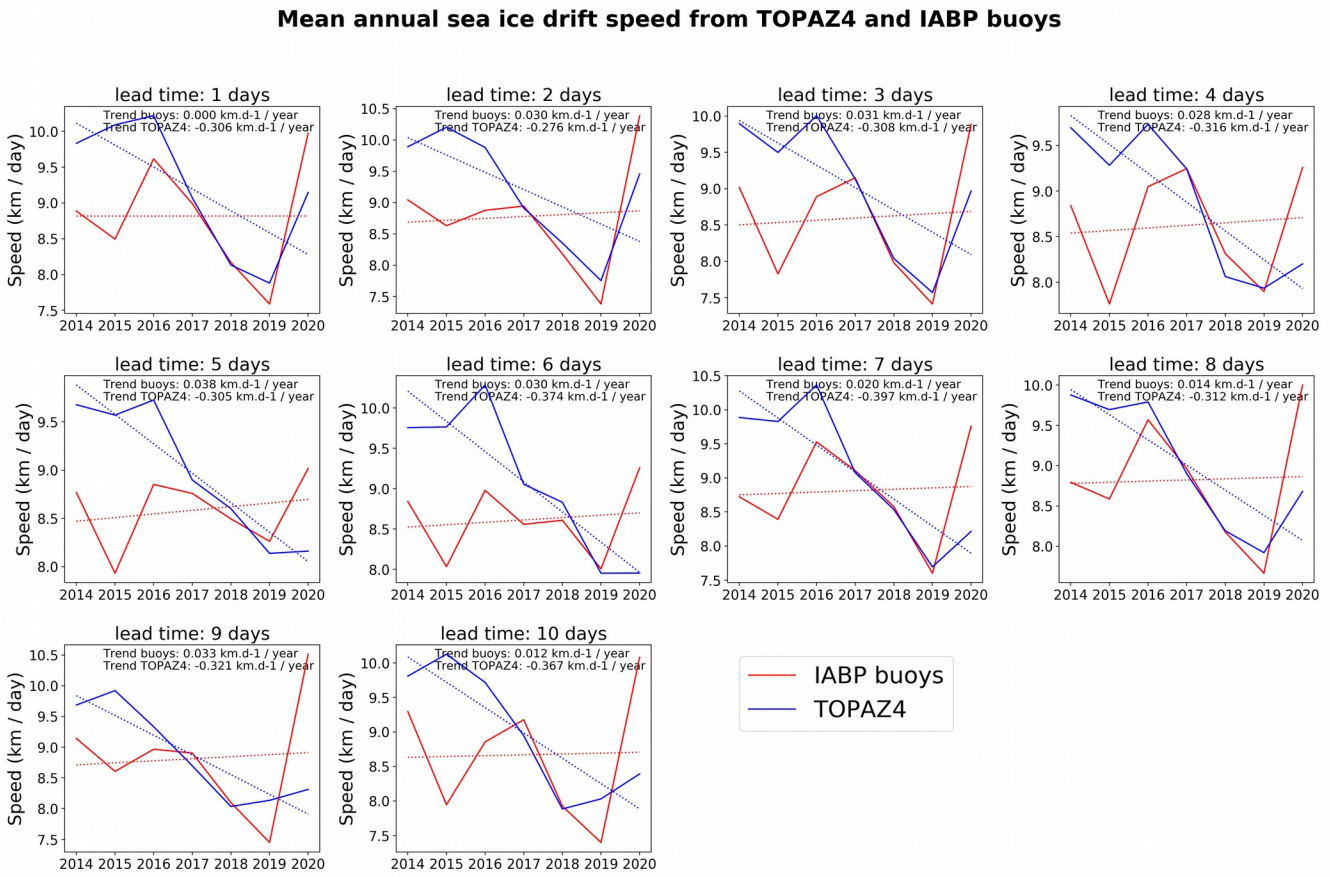


Figure S6. Mean annual sea ice drift speed from collocated IABP buoy observations and TOPAZ4 forecasts. The solid lines show the mean annual sea ice drift speed (km / day) from buoy observations and TOPAZ4 forecasts. The dashed lines show the linear trends.

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

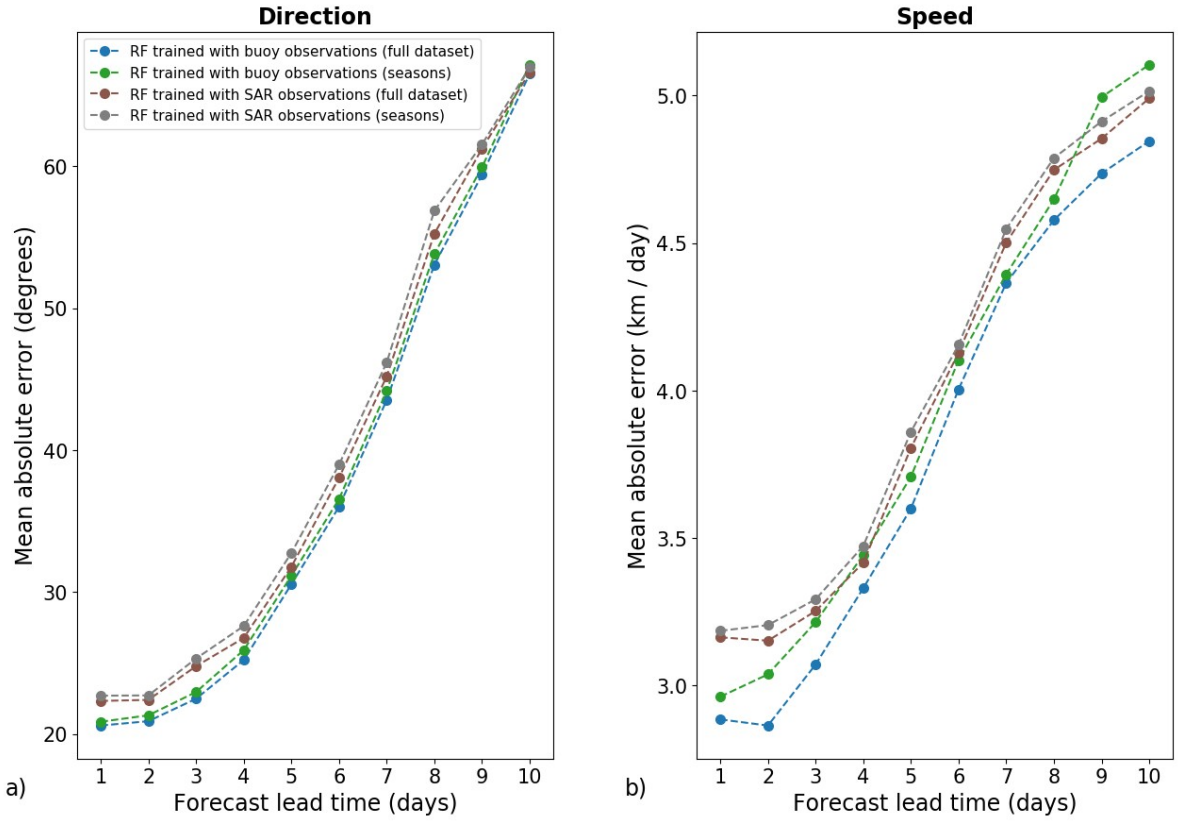
"Several training periods were tested between June 2012 and May 2020, and the chosen period from June 2013 to May 2020 seems to be optimal for predicting the direction of sea ice drift. However, using a shorter training period would have improved the forecasts for the speed of sea ice drift (figure S2 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.

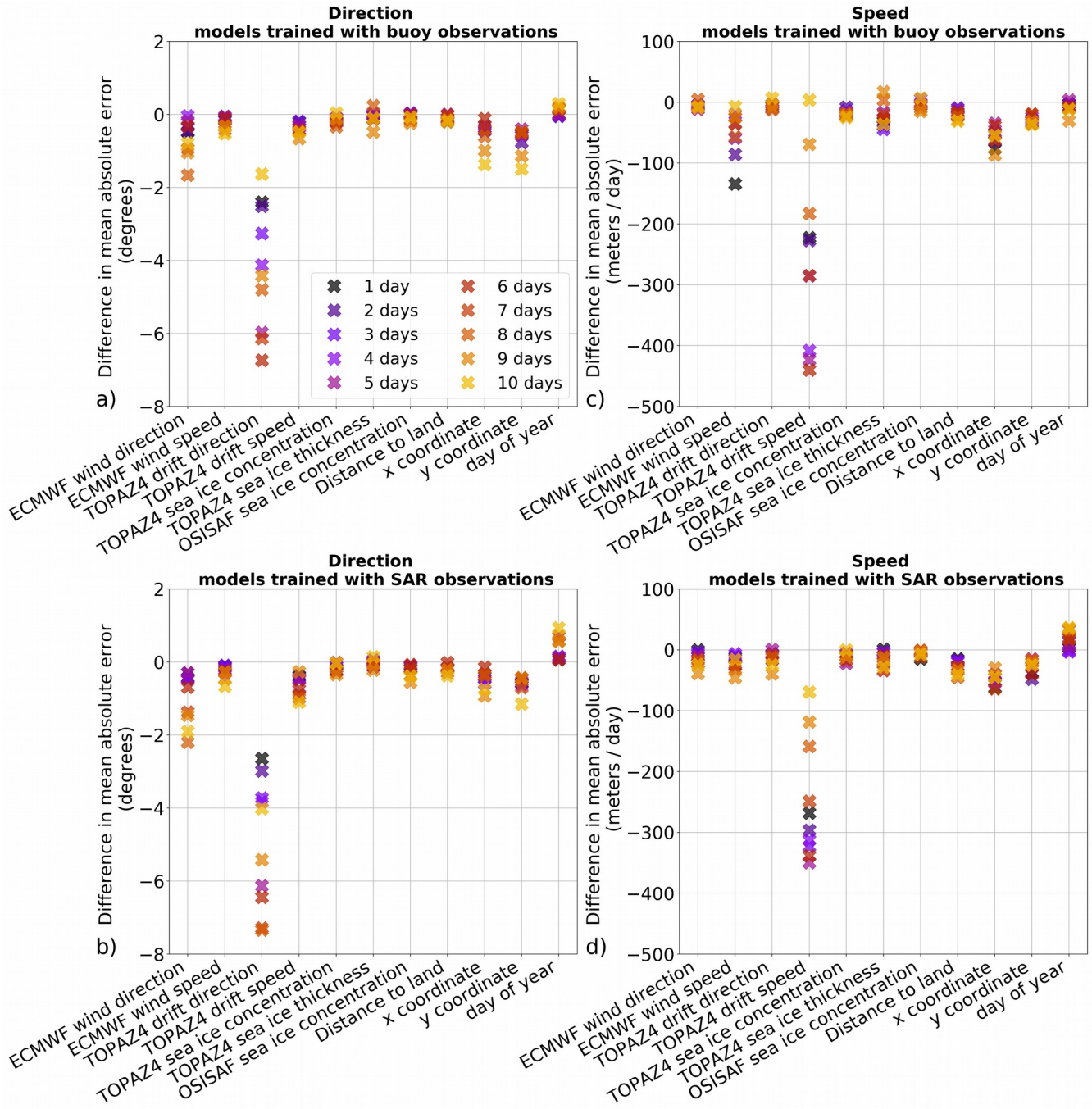


Figure S7. Differences in mean absolute error when one of the predictor variables is not used in the random forest algorithms for the direction (a, b) and speed (c, d) of sea-ice drift. The results are shown for the algorithms trained with buoy observations (a, c), and for the algorithms trained with 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):

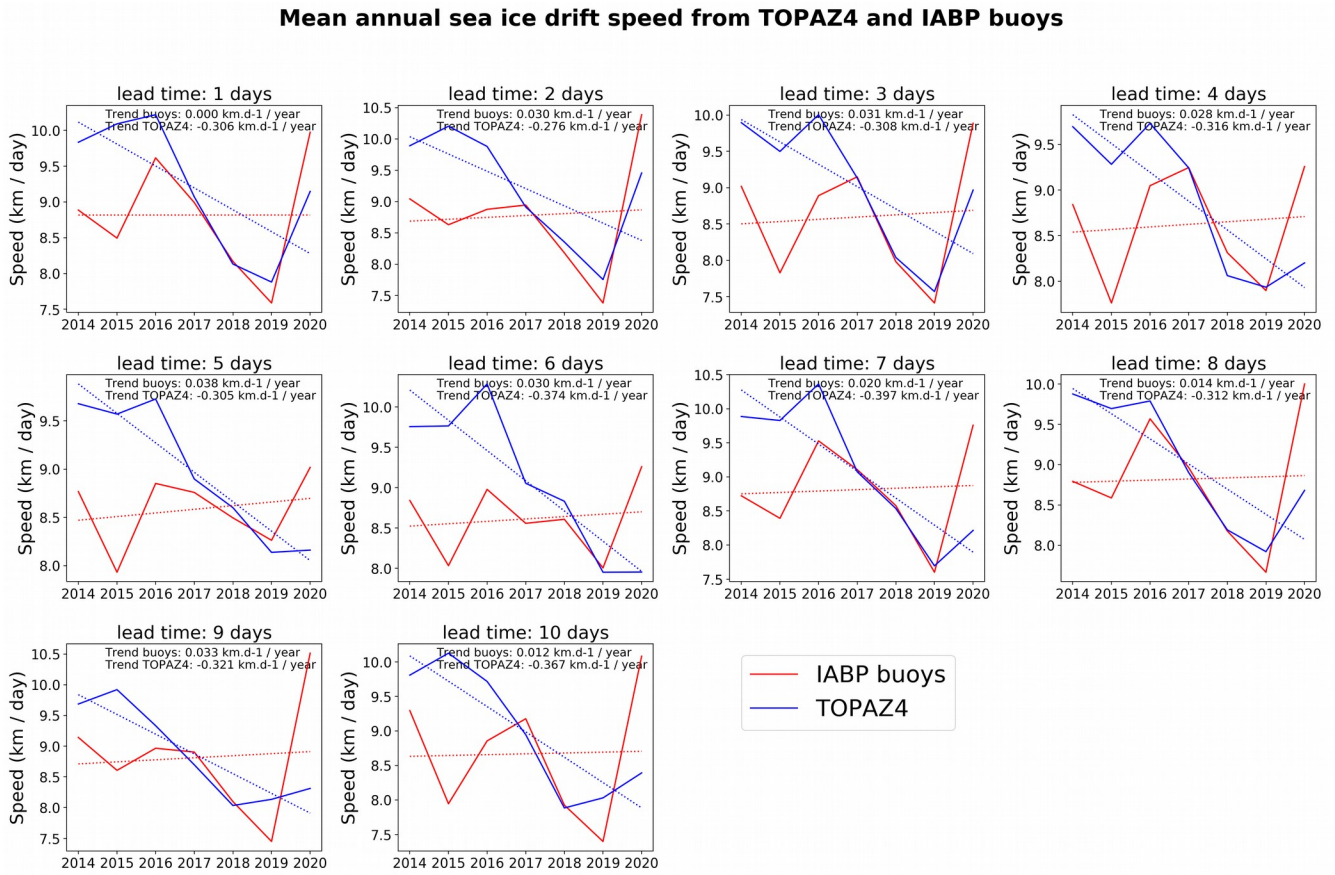


Figure S6. Mean annual sea ice drift speed from collocated IABP buoy observations and TOPAZ4 forecasts. The solid lines show the mean annual sea ice drift speed (km / day) from buoy observations and TOPAZ4 forecasts. The dashed lines show the linear trends.

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):

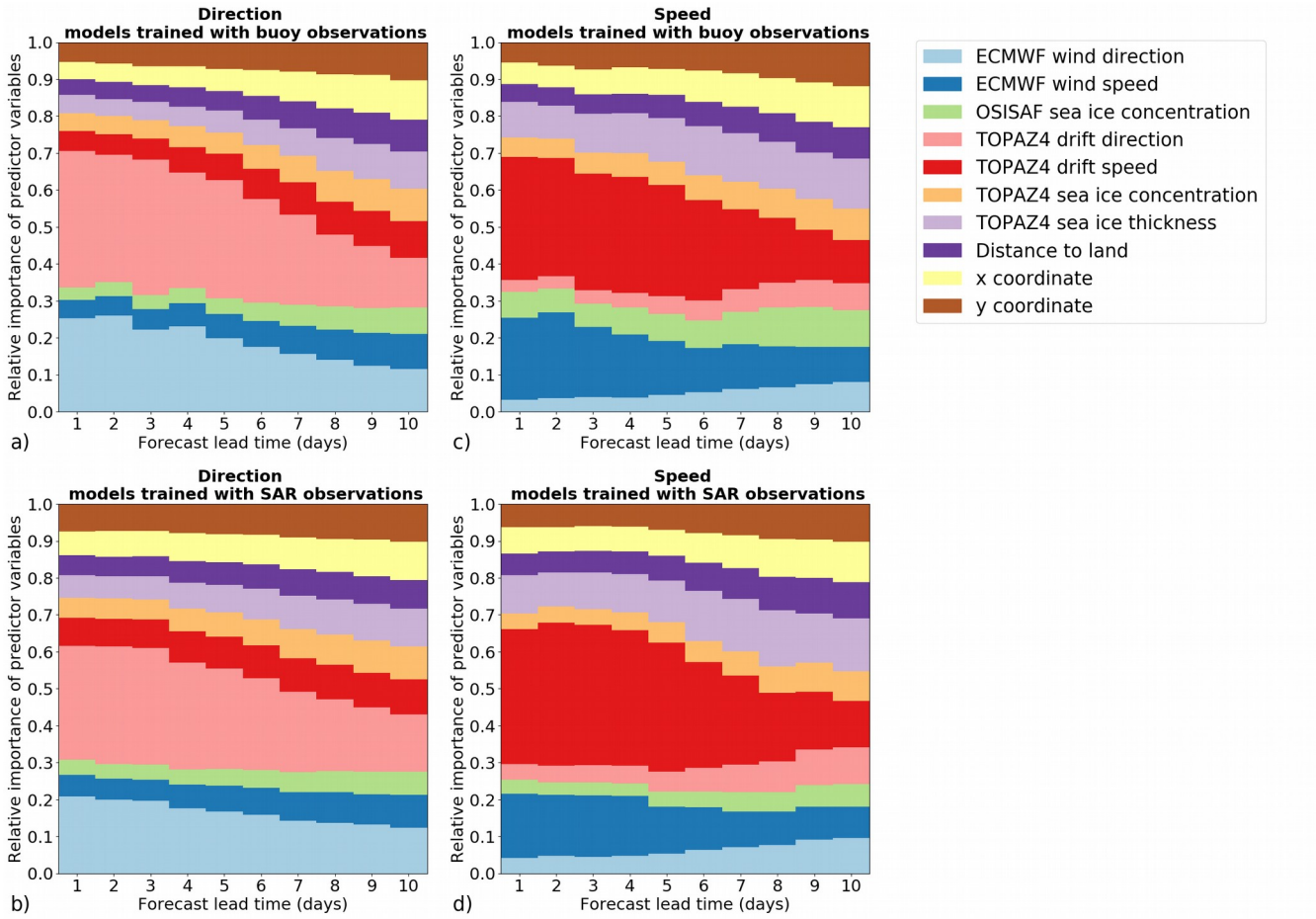


Figure 8. Relative importance of the predictor variables for the direction (a, b) and the speed (c, d) of sea ice drift assessed using the impurity-based feature importance method.

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.

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 3

##########

Title: Calibration of sea ice drift forecasts using random forest algorithms
Authors: Cyril Palerme and Malte Müller

This papers present short-term (1-10 days) forecasts of sea ice drifts using a random forest (AI) algorithm and a comparison of the AI forecasts with those of the operational ice-ocean prediction system TOPAZ. The models were trained using buoy or radarsat-derived sea-ice drifts. Predictors include short-term forecast ice speed and angle, wind speed and angle and ice thickness. Results show that the AI forecasts are more skillful than those of TOPAZ irrespective of the training data set. Furthermore, the model trained using sea ice buoys is more skillful in predicting sea ice drift for all lead-time when compared with the model trained with radarsat ice drifts.

The paper addresses an interesting question. The use of AI in sea-ice forecasting is relatively new and for this reason, this is a welcome contribution. The paper however is not well written, the introduction is succinct and does not place the work in the context of previous effectively, the model section is entirely missing and there is relatively little discussion of the pre-processing of the input data and its impact on the forecast skill (a factor that is at least equally important as the AI algorithm in producing a skillful model).

I recommend that the paper be accepted for publication after the comments below have been addressed (i.e. not rebutted).

Major Points:
1- The paper must be substantially edited/restructured.

I) The introduction is vague, there is a lot of name-dropping but it does not present an in-depth description of the previous work that is required to fully appreciate the content of the paper. I suggest that the authors review the literature more in-depth and revise the introduction substantially, or add a third co-author that works more closely in the field of sea ice forecasting.

We have completely changed the introduction in order to add a more in-depth description of previous works. We hope that the new introduction meets the expectations of the reviewer. The new introduction:

"Passive microwave observations of sea ice concentration have been available for more than 40 years, and have shown negative trends in Arctic sea ice extent since the beginning of the satellite era (e.g., Cavalieri and Parkinson, 2012; Comiso et al., 2017), with particularly strong trends during the summer (e.g., Comiso et al., 2017). There have been less satellite observations of sea ice thickness, and these retrievals have mainly been restricted to the winter due to issues related to surface melting during the summer (Ricker et al., 2017; Petty et al., 2020). Nevertheless, long-term negative trends in sea ice thickness have also been assessed by comparing retrievals from satellite altimeters (ICESat and CryoSat-2) with submarine measurements during the period 1958 - 2000 (Kwok and Rothrock, 2009;

Kwok, 2018). Furthermore, an acceleration of sea ice drift has been observed using drifting buoys and satellite observations (Rampal et al., 2009; Spreen et al., 2011; Tandon et al., 2018; Tschudi et al., 2020), and has been suggested as being a consequence of decreases in sea ice thickness and concentration due to reduced sea ice strength (Rampal et al., 2009; Olason and Notz, 2014; Tandon et al., 2018).

As a result of these changes, the Arctic ocean is becoming more accessible to marine operations, and there is an increase in maritime traffic (Eriksen and Olsen, 2018; Berkman et al., 2020). In order to ensure maritime safety, it is essential that accurate sea ice information is delivered to marine end-users. National ice services manually produce high-resolution sea ice charts using retrievals from various satellites such as passive microwave radiometers, optical instruments, and synthetic aperture radars (SAR). In addition to sea ice charts, short-term sea ice forecasts are also necessary for planning activities and providing up-to-date information to end-users. However, the spatial resolution of current sea ice models is often too coarse compared to user needs.

Short-term sea ice drift forecasts are operationally produced by numerical prediction systems, but are affected by biases despite the numerous efforts for improving the models (Hebert et al., 2015; Schweiger and Zhang, 2015; Rabatel et al., 2018;Williams et al., 2019). Hebert et al. (2015) evaluated sea ice drift speed forecasts from the U.S. Navy's Arctic Cap Now-cast/Forecast system. They found that the predicted ice drift speed was slower than drifting buoys in the summer months, and that a persistence forecast was generally better than the forecasts from the prediction system during the summer. In contrast, the forecasts produced by the U.S. Navy's Arctic Cap Nowcast/Forecast system outperformed persistence forecasts during the winter months. Schweiger and Zhang (2015) evaluated forecasts of sea ice drift speed from the Marginal Ice Zone Modeling and Assimilation System (MIZMAS) and found root mean square errors from 4.5 to 8 km per day for lead times of 1 and 9 days, respectively. These forecasts outperform a climatological reference for all lead times (up to 9 days). Sea ice drift forecasts from the neXtSIM-F system have been evaluated by Rabatel et al. (2018) and Williams et al. (2019), and root mean square errors of about 3 and 4 km per day have been reported for lead times of 1 and 4 days, respectively (Williams et al., 2019).

Sea ice drift is influenced by various sea ice characteristics such as concentration and thickness, as well as by near-surface wind and ocean currents (Rampal et al., 2009; Spreen et al., 2011; Olason and Notz, 2014; Yu et al., 2020). 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). In order to improve the accuracy of sea ice drift forecasts, we have developed two calibration methods using random forest algorithms (Breiman, 2001), which is a supervised machine learning technique suitable for assessing nonlinear relationships between a set of predictor variables and a target variable.

While random forest methods have been widely used in sea ice remote sensing (Miao et al., 2015; Han et al., 2016; Lee et al., 2016; Gegiuc et al., 2018; Park et al., 2020), as well as in weather forecasting (Gagne et al., 2014; Ahijevych et al.,2016; Herman and Schumacher, 2018; Loken et al., 2019; Mao and Sorteberg, 2020), there has been less interest in using random forests in sea ice forecasting. Recently, Kim et al. (2020) developed and compared 1-month 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). Furthermore, other machine learning and statistical methods have been used for sea ice forecasting, particularly for predicting the sea ice concentration and extent. Wang et al. (2019) used a

vector autoregressive model and a vector Markov model to predict sea ice concentration at subseasonal timescales, and obtained the best results using the vector Markov model. The vector Markov model also significantly outperformed the National Centers for Environmental Prediction (NCEP) Climate Forecast System, version 2 (NCEP CFSv2) for lead times between 2 and 6 weeks. Comeau et al. (2019) used a method based on analog forecasting for predicting Arctic sea ice area and volume anomalies at seasonal time scales, and obtained improvements compared to damped persistence forecasts. Moreover, various neural networks have been used for predicting sea-ice concentration, and found to be skillful for 1 and 12 month forecasts (Chi and Kim, 2017; Kim et al., 2020), but only slightly better than persistence forecasts for short-term prediction (Fritzner et al., 2020). Nevertheless, there has not been any attempt to calibrate short-term sea ice drift forecasts using advanced statistical methods.

The random forest models developed in this study are based on predictor variables from sea ice forecasts produced by the Copernicus Marine Environment Monitoring Service's (CMEMS) TOPAZ4 prediction system (Sakov et al., 2012), wind forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF), and sea ice satellite observations from the Ocean and Sea Ice Satellite Application Facility (OSI-SAF). While all the models use the same predictor variables, two sets of models were developed using either drifting buoy displacements or SAR observations for the target variables. The data and methods used in this study are presented in sections 2 and 3, respectively. In section 4, the daily SAR observations used for analyzing the spatial variability of the forecast errors, as well as for training some of the random forest algorithms, are evaluated using buoy observations. Then, the performances of the calibrated forecasts are evaluated and compared to those from the TOPAZ4 forecasts in section 4. The discussion and conclusion of this study are presented in section 5."

ii)The use of the present perfect to describe the data is odd. "Satellite sea-ice drift observations ... have been used..." (Line 47, 56, etc.). It sounds as though the authors are speaking of previous work by other authors when they are speaking of their own work being presented. The use of the present tense is much more engaging for the reader, or at least the simple past. I.e. We use (used) satellite sea-ice drifts observations..." . These are two examples; there are many more in the paper.

We agree with this comment and we have replaced the present perfect by the preterit many times in the revised version of the paper.

Iii) Line 67: "Section 2.2: Data used for the predictor variable". The model used as a reference for the evaluation of the AI models (i.e. TOPAZ) is included here, yet it is not a predictor variable. The forecasted sea ice thickness, concentration, drift speed and angle are all predictor variables but are not described in this section. Only the 10-m wind speed is discussed.

This section was initially about the data sets used in this study, and not about the predictor variables. However, we have decided to describe the predictor variables in this section in the revised version of the paper:
"The list of predictor variables is the same for all the models developed in this study, and can be divided into three different categories. First, some geographical information is used with the Cartesian coordinates of the grid points (x and y in the stereographic projection from the TOPAZ4 system), and the distance of the grid point to the nearest coastline in the TOPAZ4 system. Then, the sea ice concentration from passive microwave observations during the day preceding the forecast start date is also used as predictor variable. The variables from sea ice and wind forecasts during the predicted lead time can be considered as the last category. These variables are the wind direction and speed from

ECMWF forecasts, as well as the sea ice concentration, thickness, drift speed and direction from TOPAZ4 forecasts. Furthermore, the sea ice drift and concentration observations, as well as ECMWF wind forecasts, were projected onto the grid used in the TOPAZ4 prediction system using nearest-neighbor interpolation before developing the random forest models."

iv) TOPAZ is described only very succinctly. It does not say which sea ice model is used, whether there is an ice thickness distribution included, the grid on which the equation are solved, etc.

We agree with this comment and we have added a more detailed description of TOPAZ4 in section 2.2 of the revised version of the paper:
"TOPAZ4 is a coupled ice-ocean model for the North Atlantic and the Arctic which provides 10-day forecasts at a spatial resolution of 12.5 km, as well as a reanalysis (Sakov et al., 2012). It uses the version 2.2 of the Hybrid Coordinate Ocean Model (HYCOM; Bleck, 2002; Chassignet et al., 2006) coupled with a one thickness category sea ice model using an elastic-viscous-plastic rheology (Hunke and Dukowicz, 1997) derived from the version 4.1 of the Community Ice CodE (CICE). The model native grid created using conformal mapping has a spatial resolution between 12 and 16 km in the whole domain. An ensemble Kalman filter is used to assimilate satellite sea ice and oceanic observations such as sea ice concentration and drift, along-track sea-level anomalies, sea-surface temperature, as well as in-situ temperature and salinity profiles. Moreover, TOPAZ4 is forced by ECMWF high-resolution weather forecasts at the ocean surface. While TOPAZ4 forecasts are produced daily, data assimilation is only performed on Thursdays, and only the forecasts starting on Thursdays are stored in the long-term archive. Though the TOPAZ4 system provides forecasts with hourly time steps, the forecasts with daily outputs were used here due to the 24-hour span of SAR observations. Previous studies have reported that the speed of sea ice drift is overestimated in the TOPAZ4 system compared to buoy observations from the IABP (Sakov et al., 2012; Xie et al., 2017)"

2- The model is section is entirely missing. A mathematical description of the random forest model must be given because AI is relatively new in the field of short-term sea ice forecasting and more simply for the sake of completeness. The reader should not have to read other papers about random forest in order to fully appreciate the content of the current work.

We agree that the description of the random forest technique was missing, and we have added a longer description of the random forest technique in the section 3.1 of the revised version of the paper:
"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 (Pedregosaet al., 2011), and the mean squared error was used to measure the quality of the splits."

3- Some pre-processing was done to the data. E.g. the authors used speed and angle rather than latitudinal and meridional components; two different models for speed and angle were proposed. All these decisions leads to improvements in the forecast. Was there any more pre-processing done to the data to improve skill? What was the improvement in the forecast skill using these pre-processing techniques? A few sentences should be included in the discussion about this in section 4.3. I would call this section "Pre-processing of the data".

The reason why we develop different models for predicting the speed and direction of sea ice drift is that random forest algorithms can only be used to predict one variable. We hope that the new section describing the random forest technique will help to understand this better.

We have added the following statement at the beginning of the new section 3.3 called "Pre-processing of the data":

"In order to avoid overfitting, it is better to use predictor variables that are not highly correlated. This is why the speed and direction of sea-ice drift, as well as the wind speed and direction, have been used as predictor variables instead of the eastward and northward components. "

Furthermore we have added the following paragraph at the end of the same section:

"We also tested random forest models predicting the sea ice drift along the x and y axes of the TOPAZ4 grid using a different set of predictor variables (figure S12 of the supplementary material). For these models, the northward and eastward components of the ECMWF wind forecasts were used as predictors instead of the wind speed and direction, as well as the sea ice drift along the x and y axes from TOPAZ4 forecasts (which are provided by TOPAZ4 outputs) instead of the sea ice drift speed and direction. The direction and speed of sea ice drift were then calculated using the start and end location of the sea ice for comparing those models with the ones directly predicting the direction and speed of sea ice drift. Relatively similar performances were achieved by these models for predicting the direction of sea ice drift, but these models had significantly worse performances for predicting the speed of sea ice drift (larger mean absolute errors of about 12.2 % and 13.7 % on average for the models trained with buoy and SAR observations, respectively). "

4- Line 215: A model trained within the Arctic Ocean proper should not be used to predict sea-ice drift in the land-lock sea ice of the Canadian Arctic Archipelago. This is an entirely different dynamical regime. This results and associated discussion should be removed from the paper and from the abstract. Or at least not given such an important presence.

We agree with this comment, though random forest algorithms could, in principle, be able to identify regions with particular conditions with the spatial coordinates. Nevertheless, we have decided to exclude the Canadian Arctic Archipelago in the revised version of the paper, and we have added the following statement:

"The Canadian Arctic Archipelago is excluded from our study due to the different characteristics of sea-ice drift in this region (largely influenced by the presence of narrow channels and landfast ice) compared to the rest of the Arctic. Therefore, no data located in the Canadian Arctic Archipelago were used for training and evaluating the random forest models."

Minor Points:

Line 13-14: Sea ice conditions in the Arctic do not change increasingly faster because of increase in ice drift speed. Increase sea ice drift speed is one such change associated with arctic climate change, but it is not the cause. The cause is thinning of sea ice associated

with warmer air temperature, change in cloud phase and its impact on the radiative fluxes at the surface, increased ocean heat flux that interacts closely with sea ice on the shallow arctic shelves, increased storminess in the Arctic, etc

We have modified the introduction and this statement has been removed.

Line 61: Why only use sea ice drift speed lower than 5km per day? The mean speed in the Arctic Ocean is 5km /day or ~5cm/sec. It seems that a large amount of data is being ignored without acknowledging it or without providing a rationale for doing so.

It seems that there has been a misunderstanding here. In the discussion paper, it was written (lines 59-62):
"While all buoy observations located in an area with a sea-ice concentration higher than 10 % (in the OSI-SAF product described in the next section) were used for training the random forest algorithms, only the buoys with a speed between 0.5 and 100 km per day, located in an area with a sea-ice concentration higher than 10 %, and further than 50 km from the coastlines were used for verification."

Therefore, we have never excluded the buoy observations with a speed higher than 5 km / day, but only the buoy observations with a speed lower than 0.5 km / day and higher than 100 km / day. In the revised version of the paper, we have changed the threshold of 0.5 km / day to 0.1 km / day. However, we agree that the fraction of observations excluded by this selection was missing, and we have added the following statement in the revised version of the paper:

"In order to avoid inaccurate and unrealistic values, only the buoys with a speed between 0.1 and 100 km per day, located in an area with a sea ice concentration higher than 10 %, and further than 50 km from the coastlines were used for verification. While only the buoys with a speed between 0.1 and 100 km per day were used for training the random forest models predicting the direction of sea ice drift, all the buoys with a speed lower than 100 km per day were used for training the models predicting the speed of sea ice drift in order to make them able to predict very low speed. During the period from June 2013 to May 2020, about 4.5 % and 0.1 % of the buoys had a speed lower than 0.1 km per day and higher than 100 km per day, respectively. "

Line 63: "...have been projected onto the grid used in the TOPAZ4 system". This is not useful information. What grid is used in TOPAZ4? Tri-polar? Curvi-linear? Cube-sphere? I see now that this has been defined later in the paper on Line 103. The grid must be defined when it is first discussed. Is it a Cartesian grid? Or Lat/Lon?

We have added the projection in the following sentence: *"The drift vectors from buoy observations were then projected onto the polar stereographic grid used in the TOPAZ4 system."* However, we have described the other information in section 2.2 in the description of the TOPAZ4 prediction system (see our response to major point iv).

Line 79: Which ocean observations are assimilated?

We have added this information in the revised version of the paper:
"An ensemble Kalman filter is used to assimilate satellite sea ice and oceanic observations such as sea ice concentration and drift, along-track sea-level anomalies, sea-surface temperature, as well as in-situ temperature and salinity profiles."

Line 86: When did the switch to higher resolution happened?

We have changed this sentence in the revised version of the paper: *"These forecasts have lead times up to 10 days, and the model's spatial resolution changed from about 16 km to 9 km in March 2016 (https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model)."*

Line 95: No new paragraph here. "... where R is the Earth's radius, lamda and phi are the..."

We agree with this comment and we have modified this. The new sentence:
"where arctan2 represents the 4-quadrant inverse tangent function, R is the Earth's radius, φ and λ represent the latitude and the longitude, and the subscripts "start" and "end" indicate the start and end locations"

Equ 4: Unusual notation. arctan(v/u)?

It is true that "arctan2" was not defined in the text. We have added the following statement in the text: *"where arctan2 represents the 4-quadrant inverse tangent function".*

Line 121: Should it be "data points" instead of "data sets"?

We agree with this comment, and we have replaced "data sets" by "data points".

Line 165, Equ. 5: Why Case #3 in Equ. 5? Don't Case #1 and #2 above cover all cases?

There are also cases where the difference between two directions is between -180 and 180°. However, there was an error in the discussion paper (but not in the analysis). We have corrected this error in the revised version of the paper. When $\Delta D > 180$ => Error = $\Delta D - 360$ (and not $360 - \Delta D$). Note that this error only affected the direction error, but not the absolute error.

Line 169-171: This is "Method" material that was already covered earlier. It should be moved to the method section.

We have removed this statement in the revised version of the paper.

Line 191. "Moreover the fraction of forecasts improved by the calibration is, on average, larger for the models trained with buoy observations (57.0 %) than for the models trained with SAR observations (54.8 %)". Is this really statistically significant? Errors are provided throughout the paper but it does not transpire in the discussion. The errors should used to assess whether the improvements are significant or not.

We have added an analysis of the statistical significance using the Wilcoxon signed-rank test, which is suitable for non-parametric data and paired observations (see paragraph below). Note that we have used the Wilcoxon signed-rank test to assess if the difference in absolute errors are significant. The fraction of forecasts improved does not have any statistical distribution, and it is therefore more difficult to assess the statistical significance for this metric.

We have added the following paragraph in the method section:

"In this study, we used the Wilcoxon signed-rank test to assess the statistical significance of the differences between the absolute errors due to its suitability for non-parametric data (the absolute errors are not normally distributed) and paired observations (the same data set was used for evaluating the different models). We performed this analysis using the two-tailed hypothesis test and the significance level of 0.05."

And we have describe the statistical significance of the results in the section "4.2 Evaluation of the calibrated forecasts":

"The performances of the calibrated forecasts have been evaluated and compared to those from the TOPAZ4 prediction system during the period from June 2020 to May 2021 using buoy observations (figures 3). For predicting the direction of sea ice drift, the models trained with buoy observations significantly outperform the TOPAZ4 prediction system and the models trained with SAR observations for all lead times, except 10 days. On average, the calibrated forecasts produced by these models have a mean absolute error about 8.0 % lower than TOPAZ4 forecasts. The models trained with SAR observations significantly outperform the TOPAZ4 prediction system for lead times up to 5 days, and reduce the mean absolute errors by 3.3 % compared to TOPAZ4 forecasts. However, the TOPAZ4 prediction system slightly outperform the models trained with SAR observations for lead times from 8 to 10 days, though the differences are not statistically significant. Moreover, the fraction of forecasts improved by the calibration is, on average, larger for the models trained with buoy observations (55.7 %) than for the models trained with SAR observations (52.9 %). Furthermore, the correlation between the forecasts and the buoy observations is improved by both calibration methods for lead times up to 7 days, and deteriorated for longer lead times.
For the speed of sea ice drift, the models trained with buoy observations have the best performances for all lead times. They significantly outperform the TOPAZ4 system and the models trained with SAR observations for all lead times, except 4 days for which the difference with the TOPAZ4 system is not statistically significant. The forecasts from the models trained with SAR observations have slightly larger mean absolute errors than TOPAZ4 forecasts for lead times up to 5 days, but significantly outperform TOPAZ4 forecasts for longer lead times. On average, the mean absolute error is reduced by 7.1 % and 2.5 % by the calibration for the models trained with buoy and SAR observations, respectively. The fraction of forecast improved is, on average, slightly larger for the models trained with buoy observations (53.4 %) than for the models trained with SAR observations (53.1 %). Moreover, the correlation between the buoy observations and the forecasts is improved by both calibration methods.
The spatial variability of the fraction of forecasts improved by the calibration has been analyzed using SAR observations as reference in order to use as many observations as possible (figures 4, 5, 6, 7), though the grid points with less than 20 SAR observations during the period from June 2020 to May 2021 have been excluded from this analysis. The number of SAR observations per grid cell used for this comparison has been mapped in figure 1 d). Overall, both calibration methods perform relatively well for predicting the direction of sea ice drift in the Central Arctic for lead times up to 5 days (figures 4 and 5). However, the fraction of forecasts improved decreases with increasing lead times, and both calibration methods have relatively poor performances in the Beaufort, Chukchi, and East Siberian seas. Furthermore, the models trained with buoy observations perform better than the models trained with SAR observations in most of the area taken into account in this analysis.
For the speed of sea ice drift, the models trained with SAR observations perform better than the models trained with buoy observations in most of the area analyzed. The models trained with buoy observations have particularly poor performances compared to TOPAZ4 near the Greenland and Russian coastlines (figure 6), while the models trained with SAR observations perform better in these

areas (figure 7). It is worth noting that most of the buoys taken into account for evaluating the forecasts in figure 3 are not located in the areas where the models trained with buoy observations have poor performances, which likely explains the better performances of the models trained with buoy observations compared to the models trained with SAR observations in figure 3."

Line 197: "The fraction of forecast improved is, on average, slightly larger for the models trained with SAR observations (55.3 %) than for the models trained with buoy observations (54.9 %). " Again, is this statistically significant?

We have answered to this comment in our previous response.

Line 222: The fraction of data used in the training and validation of the model belongs to the Method section.

We have moved this statement in the Method section in the revised version of the paper.

Line 225-230: Repetitive. This was already mentioned in the Method section.

We have moved this section in the Method section in the revised version of the paper.

Line 236: Sea ice thickness does not change very much in 10 days. I suspect the ice thickness at t=0 would be equally skillful. This should be mentioned.

We agree that using sea ice thickness during the initialization of the forecasts should provide a relatively similar information to the algorithms. However, because Pan-Arctic sea ice thickness observations are not available during the summer, it is not possible to use sea ice thickness observations in our random forest models which are used all year round. Therefore, we consider that the best available information is the sea ice thickness forecasts at the predicted lead time, and we do not think that the low temporal variability of sea ice thickness should be mentioned here.

Section 4.3: The discussion does not present a quantitative assessment of the predictive skill of each predictor. A more quantitative discussion should be provided.

We agree with this comment, and we have modified this section to present the results more quantitatively. The new section:
"For both calibration methods, the most important variable for predicting the drift direction is the sea ice drift direction from TOPAZ4 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. On average, the mean absolute errors are reduced by all predictors for the direction and speed of sea ice drift in both calibration methods (figure 9), though some predictor variables do not improve the forecast accuracy for all lead times. While the sea ice concentration observations during the initialization of the forecasts and the sea ice concentration forecasts from TOPAZ4 are correlated, removing one of these variables decreases the accuracy of most random forest models. Therefore, we decided to keep both variables, even if the importances of these variables are probably underestimated due to this correlation. 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.
For the models predicting the direction of sea ice drift, removing the drift direction from TOPAZ4 forecasts increases the mean absolute error between 1.1 and 6.7 degrees depending on the lead time and the observations used for the target variable. This is much larger than the differences in mean absolute error when the wind direction from ECMWF forecasts is removed (between 0.1 and 2.2 degrees). For the models predicting the speed of sea ice drift, removing the drift speed from TOPAZ4 forecasts increases the mean absolute error between 0.041 and 0.444 km / day depending on the lead time and the observations used for the target variable. This is also much larger than the differences in mean absolute error when the wind speed from ECMWF forecasts is removed. Surprisingly, removing the wind speed forecasts slightly reduces the mean absolute error (difference of 0.005 km / day) for the model predicting the speed of sea ice drift for a lead time of 4 days trained with SAR observations. For the other models predicting the speed of sea ice drift, removing the wind speed forecasts increases the mean absolute error between 0.001 and 0.127 km / day. 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).”

Figure 1: Colorbar for the d panel should be changed to avoid saturation.

The colorbar has been changed.

Figure 4: Units for sea ice drift should be km/day or ideally cm/sec. It should not be m/day.

We have changed the unit for sea ice drift speed in the revised version of the paper, and km / day is now used.

Bruno Tremblay
McGill University