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?

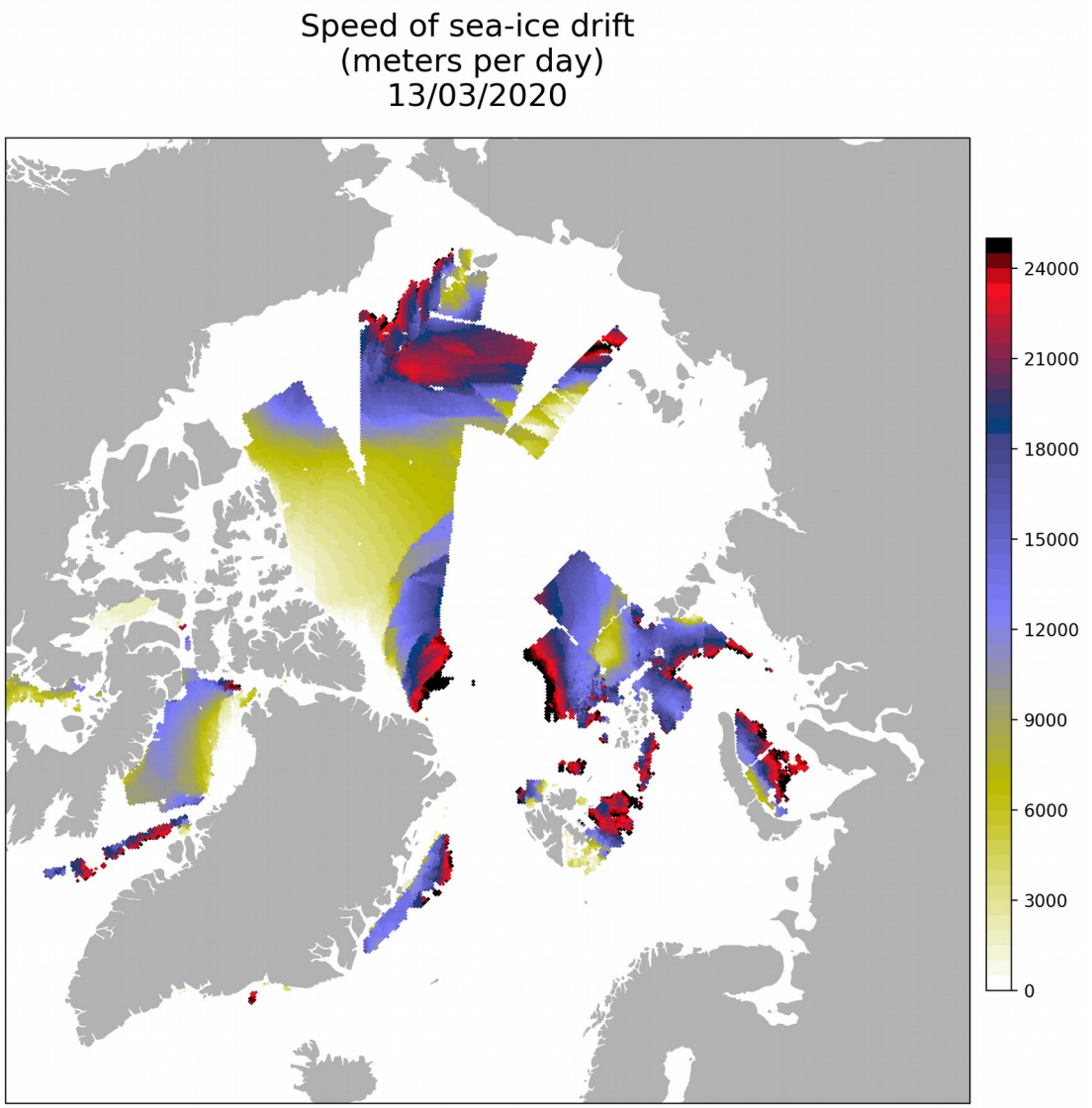
Thanks for this comment. We have checked this, and *“initial great-circle course angle”* seems to be the most common term. We have used this term in the revised version of the paper.

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"

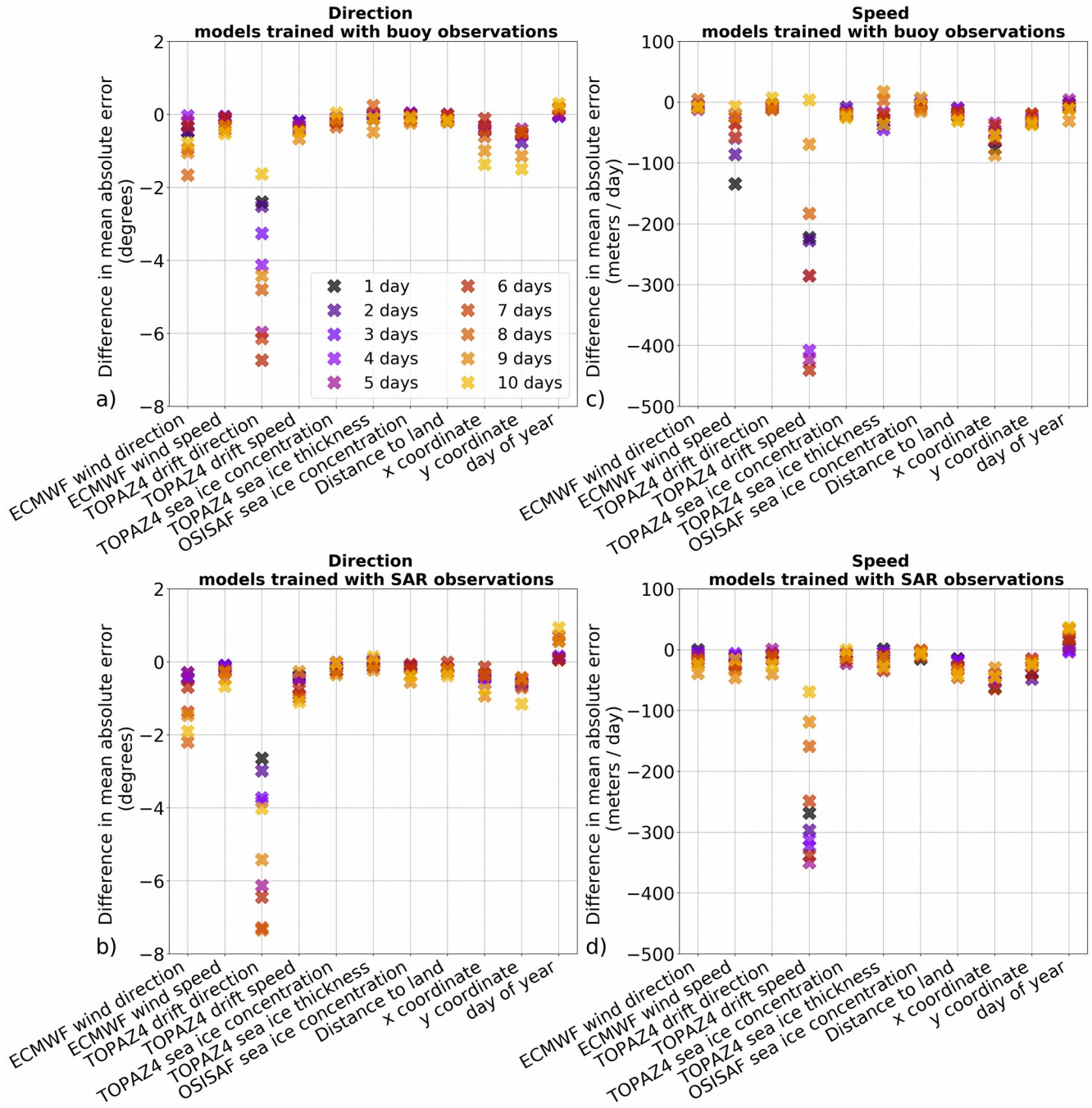


Figure S7. Differences in mean absolute error when one of the predictor variables is not used in the random forest algorithms for the direction (a, b) and speed (c, d) of sea-ice drift. The results are shown for the algorithms trained with buoy observations (a, c), and for the algorithms trained with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

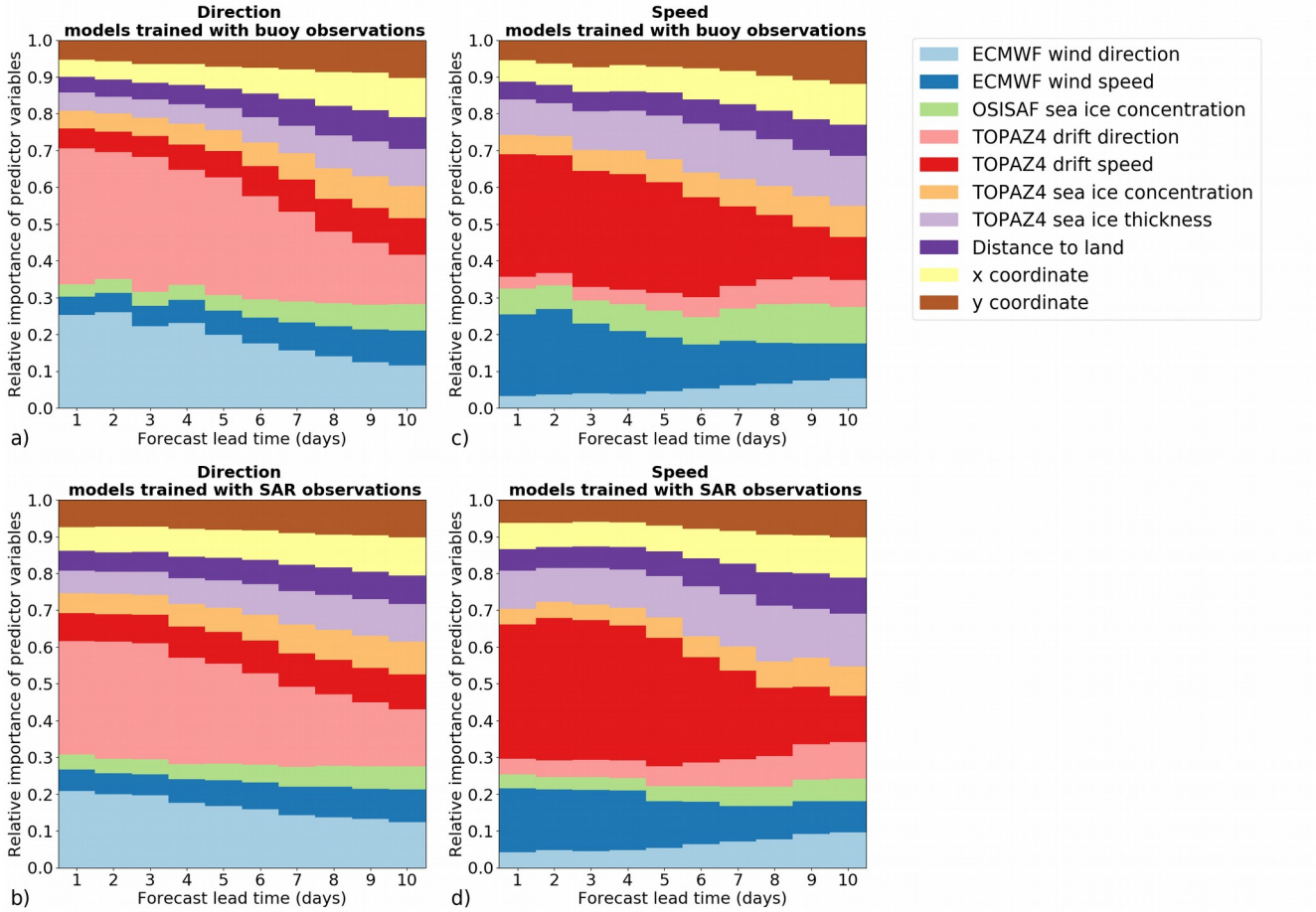


Figure 8. Relative importance of the predictor variables for the direction (a, b) and the speed (c, d) of sea ice drift assessed using the impurity-based feature importance method.

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.

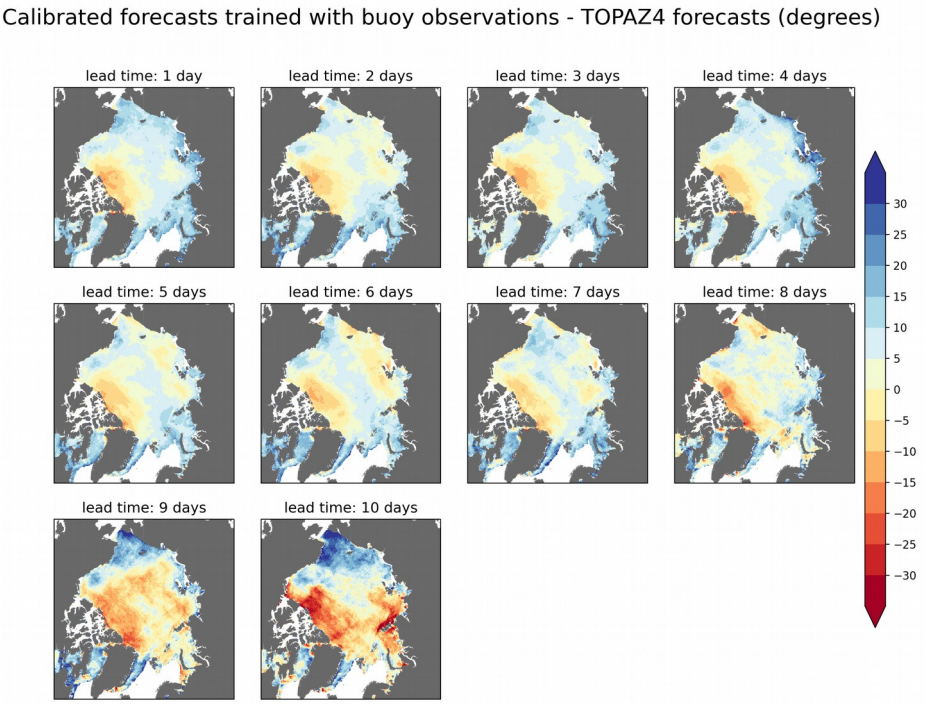


Figure S13. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

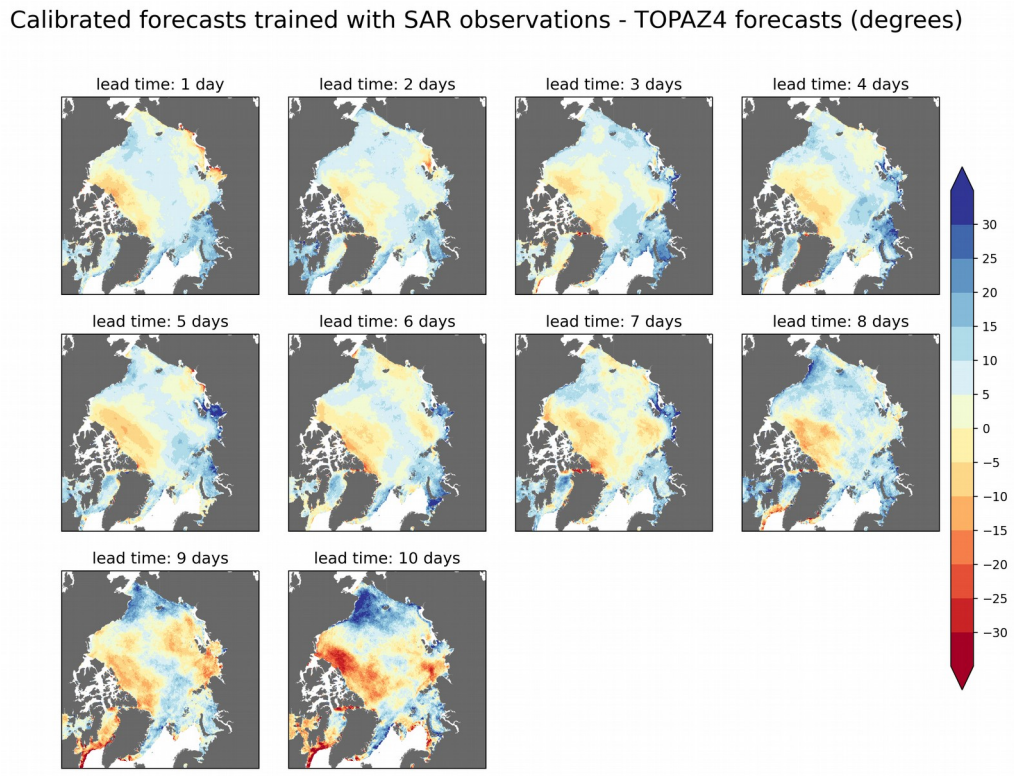


Figure S14. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the direction of sea ice drift (degrees) during the period June 2020 - May 2021.

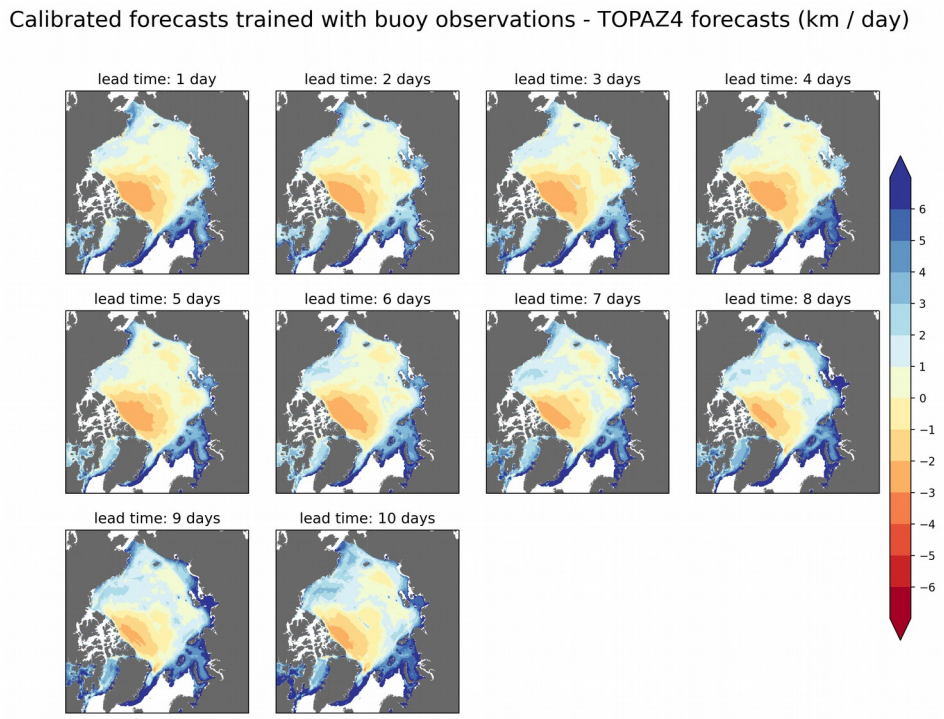


Figure S15. Difference between the random forest models trained with buoy observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

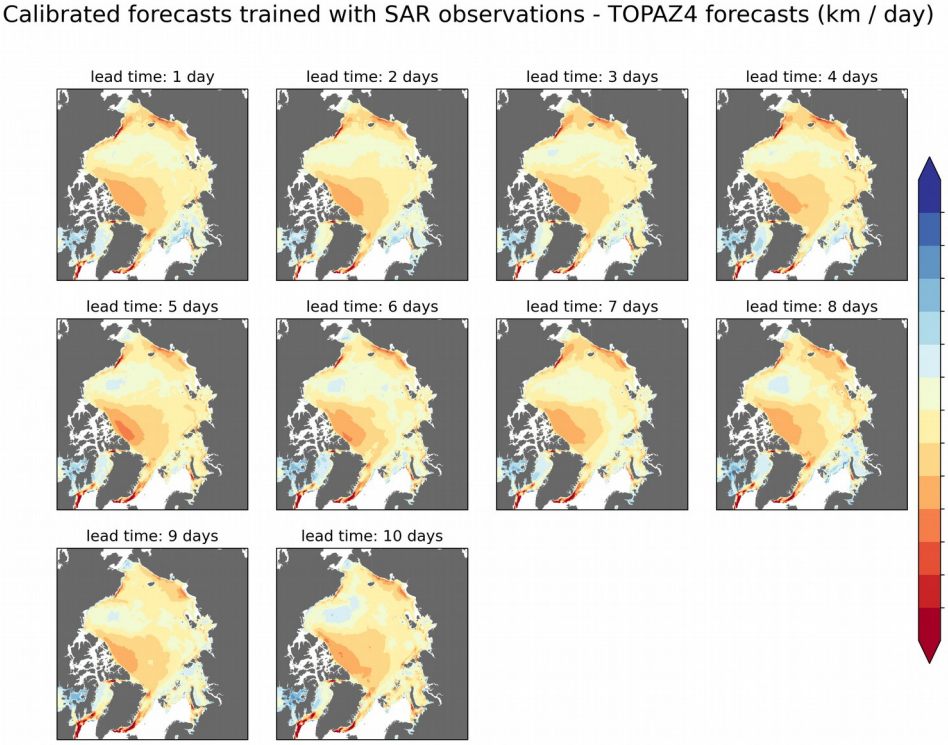


Figure S16. Difference between the random forest models trained with SAR observations and the TOPAZ4 forecasts for the speed of sea ice drift (km / day) during the period June 2020 - May 2021.

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:

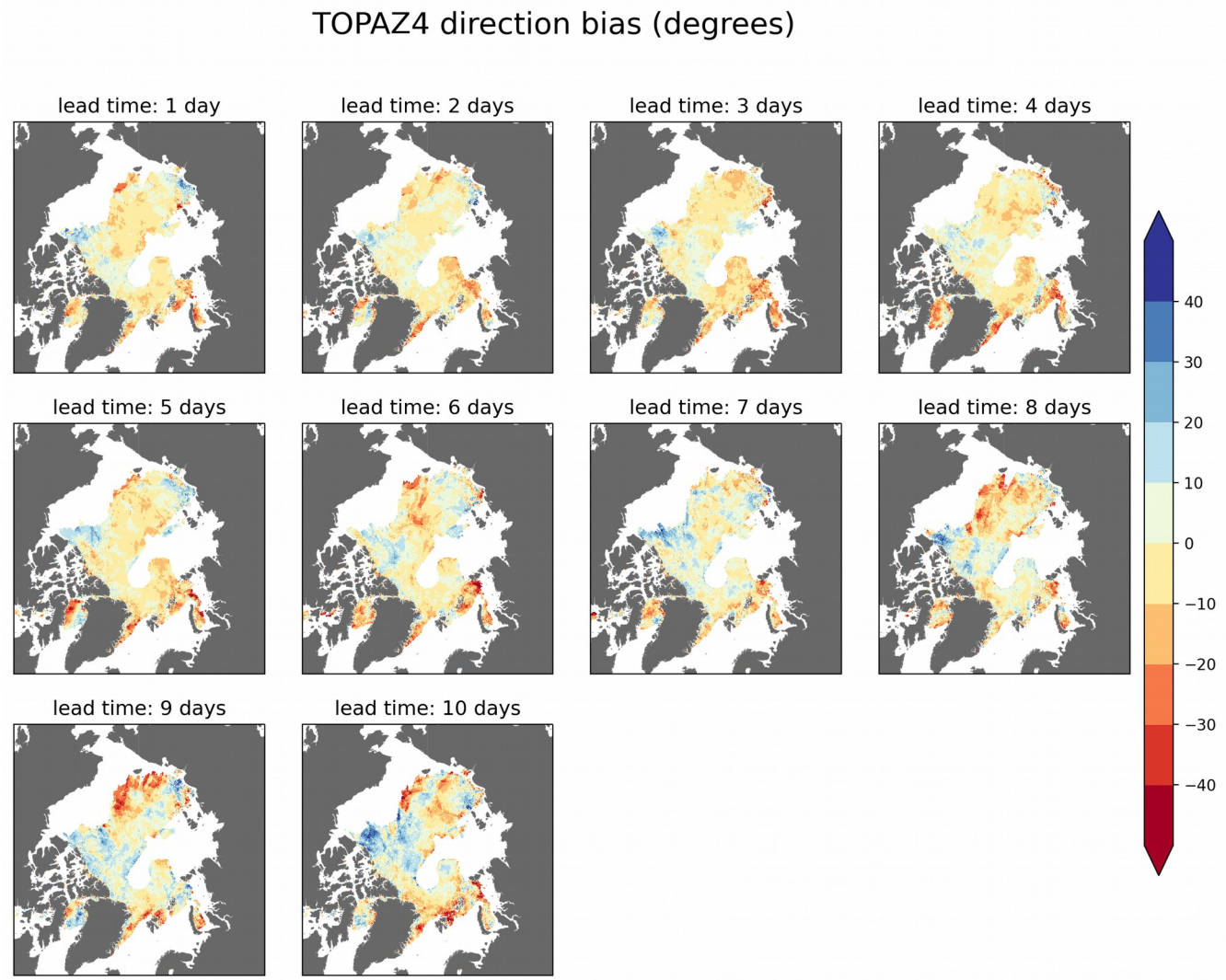


Figure S9. Mean direction bias (degrees) from TOPAZ4 forecasts during the period 2018 - 2019 compared to SAR observations. Only the grid points containing at least 20 SAR observations during the period 2018 - 2019 have been taken into account.

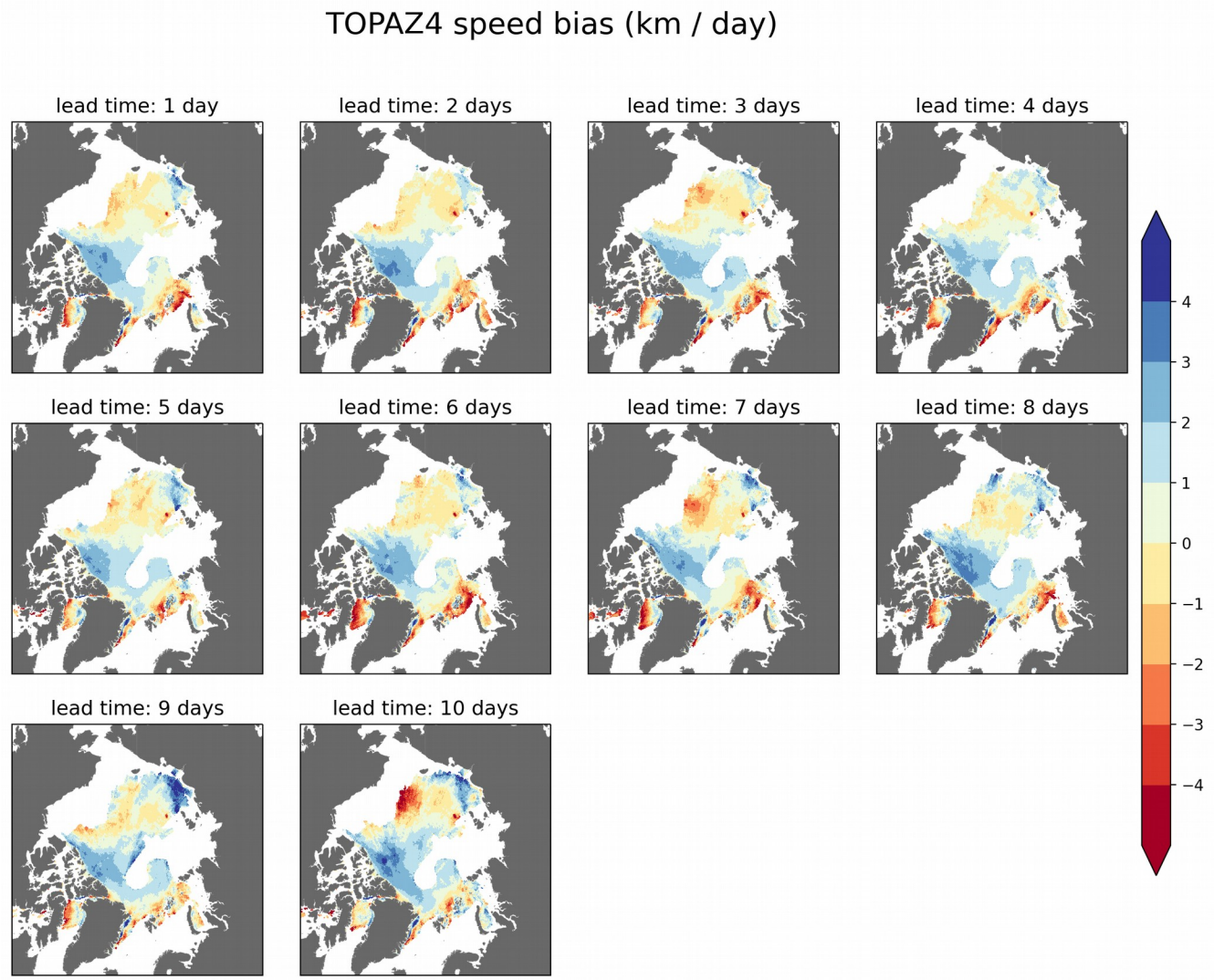


Figure S10. Mean speed bias (km / day) from TOPAZ4 forecasts during the period 2018 - 2019 compared to SAR observations. Only the grid points containing at least 20 SAR observations during the period 2018 - 2019 have been taken into account.

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.

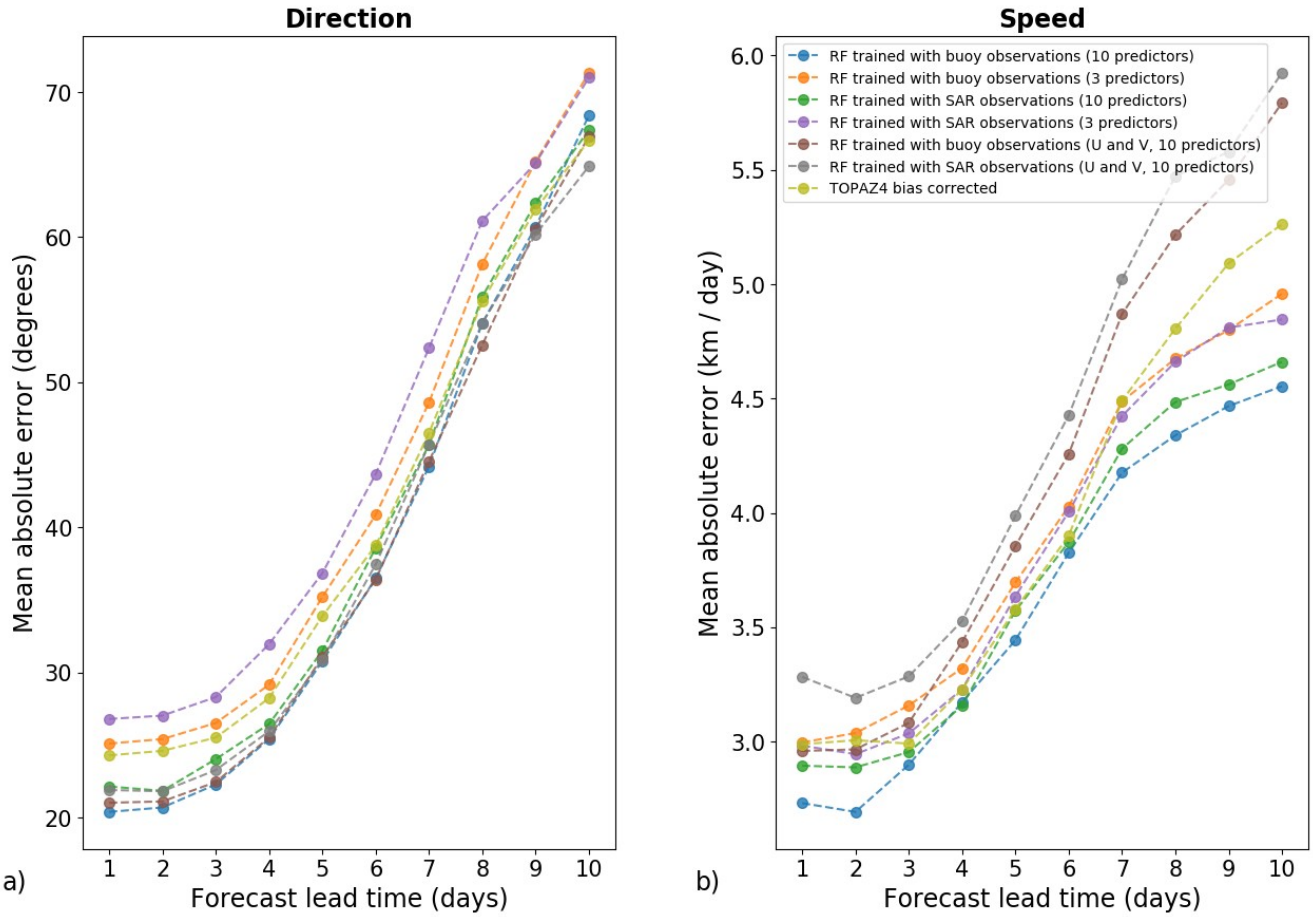


Figure S12. Mean absolute errors of different calibration methods for the period June 2020 – May 2021. Buoy observations have been used as reference. The random forest (RF) models using all the predictors (10 predictors) described in the main paper are shown by the blue and green curves. The random forest 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.