Calibration of sea ice drift forecasts using random forest algorithms

Cyril Palerme[1] and Malte Müller[1]

[1]Development Centre for Weather Forecasting, Norwegian Meteorological Institute, Oslo, Norway

Correspondence: Cyril Palerme (cyril.palerme@met.no)

Abstract. ~~Developing accurate sea-ice~~

Developing accurate sea ice drift forecasts is essential to support decision making of maritime end-users operating in the Arctic. In this study, two calibration methods have been developed for improving 10-day ~~sea-ice~~ sea ice drift forecasts from an operational ~~sea-ice~~ sea ice prediction system (TOPAZ4). The methods are based on random forest ~~algorithms~~ models (supervised machine learning~~models) and have been trained using either~~) which were trained using target variables either from drifting buoy or synthetic-aperture radar ~~observations for the target variables~~(SAR) observations. Depending on the calibration method, the mean absolute error is reduced, on average, between ~~5.9 % and 8.1~~ 3.3 % and 8.0 % for the direction, and between ~~7.1 % and 9.6~~ 2.5 % and 7.1 % for the speed of ~~sea-ice~~ sea ice drift. Overall, the algorithms trained with buoy observations have the best performances when the forecasts are evaluated using drifting buoys as reference. However, there is a large spatial variability in these results, and the ~~algorithms~~ models trained with buoy observations have particularly poor performances for predicting the speed of ~~sea-ice drift in the Canadian Archipelago, along the east coast of Greenland , and north of Svalbard. In these areas, the algorithms~~ sea ice drift near the Greenland and Russian coastlines compared to the models trained with SAR observations~~have better performances for predicting the speed of sea-ice drift.~~

.~.

1 Introduction

~~Due to increasing maritime traffic in the Arctic (Eriksen and Olsen, 2018; Berkman et al., 2020), there is a growing demand for reliable sea-ice information. Nevertheless, sea ice conditions in the Arctic tend to change increasingly faster due to the ongoing acceleration of sea-ice drift~~ 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, 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)~~. Therefore, accurate sea-ice drift forecasts are essential for predicting future sea-ice conditions, and for~~, 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 ~~navigating in ice-covered waters who need~~. 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 ~~(Jeuring and Knol-Kauffman, 2019)~~ to end-users. However, the spatial resolution of current sea ice models is often too coarse compared to user needs.

~~Sea-ice~~ 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)~~ (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 Nowcast/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~~ 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). ~~The~~ 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~~ 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~~ predictors and a target variable.

While random forest methods have been widely used in ~~sea-ice~~ 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~~ sea ice forecasting. Recently, Kim et al. (2020) developed ~~sea-ice~~ and compared 1-month sea ice concentration forecasts based on random forests and convolutional neural networks ~~, but they~~. They obtained more accurate results using convolutional neural networks ~~. However~~ 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~~ sea ice forecasting, particularly for predicting the ~~sea-ice~~ sea ice concentration and extent ~~(Wang et al., 2016; Chi and Kim, 2017; Comeau et al., 2019; Wang et al., 2019; Fritzner et al., 2020)~~. Wang et al. (2019) used a vector autoregressive model and a vector Markov model to predict sea ice concentration at subseasonal time scales, 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 ~~algorithms~~ random forest models developed in this study are based on predictor variables from ~~sea-ice~~ 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~~ sea ice satellite observations from the Ocean and Sea Ice Satellite Application Facility (OSI-SAF). While all the ~~algorithms~~ models use the same predictor variables, two sets of ~~algorithms have been~~ models were developed using either drifting buoy displacements or ~~synthetic-aperture radar (SAR)~~ SAR observations for the target variables. ~~Furthermore, the evaluation of the forecasts has been performed using drifting buoy displacements for assessing the forecast errors, and using SAR observations for analyzing the spatial variability of the forecast performances.~~

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.

2 Data

2.1 ~~Sea-ice~~ Sea ice drift observations

~~Satellite sea-ice~~ In this study, satellite sea ice drift observations from the CMEMS product named SEAICE_GLO_SEAICE_L4_NRT_OBSI (MOSAIC version 2.0, hereafter referred as CMEMS SAR MOSAIC product) ~~have been~~ 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~~ sea ice drift forecasts. This product provides ~~sea-ice~~ sea ice drift fields derived from SAR observations acquired by the Sentinel-1 satellites with a spatial resolution of 10 km and a temporal resolution of 24 hours. Only a fraction of the Arctic is covered by this product every day, and there is no observations north of 87.7°N (figure 1). It is worth noting that this product is an average of drift vectors derived from pairs of SAR images which are not necessarily acquired around midnight, and that the averaging process

introduces uncertainties in the product. However, this product has the advantage of providing gridded sea ice drift fields with fixed time spans (from midnight to midnight the next day) which are necessary for training random forest algorithms. Moreover, the fixed time spans ease the comparison between SAR observations and sea ice drift forecasts with daily time steps.

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. For consistent comparisons with the SAR observations and the sea ice drift forecasts, the speed and direction of sea ice drift were calculated using the geographical coordinates of the buoys at midnight (UTC). The drift vectors from buoy observations were then projected onto the polar stereographic grid used in the TOPAZ4 system. When several buoys were located in the same grid cell, only the nearest one from the grid point was taken into account. 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. The number of buoy observations used for evaluating the forecasts during the period from June 2020 to May 2021 varies between 19276 and 19576 depending on the lead time, and has been mapped in figure 1 b).

2.2 Predictor variables

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.

For the sea ice concentration observations during the day preceding the forecast start, the version 2 of the global sea ice concentration climate data record from OSI-SAF (Lavergne et al., 2019) ~~has been~~ was used for training the algorithms. This

130   dataset has a spatial resolution of 25 km and is available with a latency of 16 days. Therefore, it cannot be used for producing operational forecasts. Since June 2020, calibrated forecasts have been produced daily, and near-real-time ~~sea-ice~~ sea ice concentration products at 25-km resolution processed at the Norwegian Meteorological Institute from AMSR2 and SSMIS DMSP-F16, F17, and F18 sensors based on the algorithms introduced in Lavergne et al. (2019) have been used. The AMSR2 observations have been used when they were available, and have been replaced by SSMIS observations when they

135   were missing.

~~Short-term sea ice forecasts from the TOPAZ4 prediction system (Sakov et al., 2012) have been used in the calibration methods developed in this study.~~ TOPAZ4 is a coupled ice-ocean model for the ~~Arctic ocean~~ 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 ~~an~~ the version 2.2 of the Hybrid Coordinate Ocean Model (HYCOM; Bleck, 2002; Chassignet et al., 2006) coupled with a one

140   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 ~~. However, while~~ 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

145   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 ~~initialized using data assimilation and~~ stored in the long-term archive. Though the TOPAZ4 system provides forecasts with hourly time steps, the forecasts with daily outputs ~~have been~~ were used here due to the 24-hour span of SAR observations. Previous studies have reported that the speed of ~~sea-ice~~ sea ice drift is overestimated in the TOPAZ4 system compared to buoy observations from the ~~IAPB~~ IABP (Sakov et al., 2012; Xie

150   et al., 2017).

In addition to the OSI-SAF observations and TOPAZ4 forecasts, 10-meter wind forecasts from ECMWF are also used for the predictor variables in the random forest algorithms. These forecasts have lead times up to 10 days, and ~~during the studied period (2013 - 2020),~~ the model's spatial resolution ~~has increased~~ changed from about 16 km to 9 km in March 2016 (https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model).

155   **3   Methods**

~~In this study, the speed of sea-ice drift was determined from the great-circle distance calculated using the Haversine formula (equation 1)between the start and end locations during 24 hours. The initial bearing on the great-circle path between the start and end locations was used for the direction~~ Development ~~of sea-ice drift, and was calculated using equation 2.~~

$$D = 2R \arcsin\left(\cdot\sqrt{\sin^2\left(\frac{\varphi_{end} - \varphi_{start}}{2}\right) + \cos(\varphi_{start})\cos(\varphi_{end})\sin^2\left(\frac{\lambda_{end} - \lambda_{start}}{2}\right)}\right)$$

random forest models]Data preparation

In this study, the speed of sea-ice drift was determined from the great-circle distance calculated using the Haversine formula (equation 1)between the start and end locations during 24 hours. The initial bearing on the great-circle path between the start and end locations was used for the direction Development of sea-ice drift, and was calculated using equation 2.

$$D = 2R \arcsin\left(\cdot\sqrt{\sin^2\left(\frac{\varphi_{end} - \varphi_{start}}{2}\right) + \cos(\varphi_{start})\cos(\varphi_{end})\sin^2\left(\frac{\lambda_{end} - \lambda_{start}}{2}\right)}\right)$$

random forest models

$$\theta = \arctan 2\Big(\sin(\lambda_{end} - \lambda_{start}).\cos(\varphi_{end}),\ \cos(\varphi_{start}).\sin(\varphi_{end}) - \sin(\varphi_{start}).\cos(\varphi_{end}).\cos(\lambda_{end} - \lambda_{start})\Big)\left(\frac{180}{\pi}\right)$$

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 equations 1 and 2, 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. Furthermore, the wind speed and direction from ECMWF forecasts were calculated using equations 3 and 4, respectively, from the mean daily u and v components ($\overline{u}$ 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 $\overline{v}$ in equations 3 and 4). The computed wind direction is not the classic meteorological wind direction (direction from which the wind is blowing), but the opposite (direction the wind is blowing to) in order to be consistent with the direction of sea-ice drift calculated using equation 2.

$$WS = \sqrt{\overline{u}^2 + \overline{v}^2}$$

$$WD = \arctan 2(\overline{u}, \overline{v})\left(\frac{180}{\pi}\right)$$

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 ~~sea-ice drift speed and direction from SAR observations, the OSI-SAF sea-ice concentration observations, as well as the ECMWF wind forecasts have been projected onto the polar stereographic grid used in the TOPAZ4 system using nearest-neighbor interpolation~~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). 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.

3.1 ~~Development of statistical models~~

~~In order to predict independent variables, it has been chosen to forecast the direction~~ 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 ~~speed of sea-ice drift rather than~~ 360°. When averaging several predictions close to the ~~eastward and northward components. In this study, two sets of algorithms have been developed using target variables either from buoy displacements or from SAR observations. Moreover, different algorithms have been developed~~ 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. 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 ~~and speed~~ 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.

215   The Canadian Arctic Archipelago is excluded from our study due to the different characteristics of sea-ice drift ~~, as well as for different lead times (1 to 10 days)~~ in this region (largely influenced by the presence of narrow channels and landfast ice) compared to the rest of the Arctic. Therefore, ~~20 different models were developed using buoy displacements, and 20 other models were developed using SAR observations.~~ no data located in the Canadian Arctic Archipelago were used for training and evaluating the random forest models. Furthermore, the calibrated forecasts have been produced where the sea ice concentration
220   predicted in the TOPAZ4 forecasts was larger than 10 %.

3.1   Training data sets

Only the TOPAZ4 forecasts starting on Thursdays are stored in the long-term archive, and the algorithms have therefore been trained using weekly data. The period from January 2018 to May 2020 was used for training the algorithms with SAR observations (the CMEMS SAR MOSAIC product has been available since January 2018). Due to the smaller number of
225   available observations, a longer period (June 2013 to May 2020) was used for training the algorithms with buoy observations. While using a longer period increases the size of the training data sets, it can also introduce inconsistencies in the training data sets due to the constant development of the prediction systems (TOPAZ4 and ECMWF Integrated Forecasting System). ~~In order to optimize the performances of the algorithms trained with buoy observations, several training periods have been tested (~~Several training periods were tested between June 2012 and May 2020~~)~~, and the ~~best results have been found using the~~
230   chosen period from June 2013 to May ~~2020.~~ 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). Nevertheless, we decided to use the same training period for the random forest models
235   predicting the direction and speed of sea ice drift for consistency.

For the algorithms trained with buoy observations, the data from all the grid points where buoy observations were available within the TOPAZ4 domain have been used ~~as independent data sets~~ in order to create a large database, which results in about ~~1.7~~ 1.5 x 10⁴ ~~training data sets~~ data points for each model during the training period. However, a different approach ~~has been was~~ used for the algorithms trained with SAR observations. Taking into account all the grid points where SAR observations
240   were available in the TOPAZ4 domain would result in using many highly correlated points for training the algorithms, which increases the probability of overfitting~~(meaning that the models learn from noise in the training data)~~. In order to minimize this issue, but also taking into account a sufficient number of grid points, sensitivity tests were performed. For each forecast of the training period (January 2018 - May 2020), a random selection without replacement of the grid points has been performed, and the selected grid points were added to the training data set (similar approaches were used by Gagne et al. (2014) and Loken et al.
245   (2019) for calibrating precipitation forecasts). Random selections between 0.1 % and 100 % of the available grid points were tested, and the differences in mean absolute errors between the algorithms trained using all the available data and the algorithms trained using only a fraction of the available data were evaluated ~~using buoy observations as reference during the period from June 2020 to November 2020 (figure ??). Overall, the best results were found by selecting~~ (figure S1 of the supplementary

material). Due to the good performances of the models trained using 2 % of the ~~grid points with SAR observations. Therefore,~~ available grid points, we decided to keep 2 % of the available grid points ~~were used for the algorithms trained~~ for training the models with SAR observations~~in this study, which results in about 5.7~~. This results in using data from about 5.5 x 10^4 ~~training data sets~~ data points on average (between ~~5.5~~ 5.2 x 10^4 and ~~6.0~~ 5.7 x 10^4 data ~~sets~~ points depending on lead time). Moreover, decreasing the size of the training data sets reduces the computational cost of the algorithms. Furthermore, it is worth noting that the Arctic is not uniformly covered by SAR and buoy observations (figure 1), and that different regions have therefore different weights in the development of the algorithms.

~~The list of predictor variables is the same for the models predicting the direction and speed~~

3.2 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~~. The predictor variables can be divided into three different categories. First, some geographical information has been used with the geographical 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 startdate has also been used as predictor variable. The variables from sea-ice and wind forecasts can be considered as the last category. These variables are the wind direction and speed~~, as well as the wind speed and direction, have been used as predictor variables instead of the eastward and northward components. The speed of sea ice drift was determined from the great-circle distance calculated using the Haversine formula (equation 1) between the start and end locations during 24 hours. The initial great-circle course angle was used for the direction of sea ice drift, and was calculated using equation 2:

$$D = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_{end} - \varphi_{start}}{2}\right) + \cos(\varphi_{start})\cos(\varphi_{end})\sin^2\left(\frac{\lambda_{end} - \lambda_{start}}{2}\right)}\right) \tag{1}$$

$$\theta = \arctan 2\Big(\sin(\lambda_{end} - \lambda_{start}).\cos(\varphi_{end}),\ \cos(\varphi_{start}).\sin(\varphi_{end}) - \sin(\varphi_{start}).\cos(\varphi_{end}).\cos(\lambda_{end} - \lambda_{start})\Big)\left(\frac{180}{\pi}\right)$$
$$\tag{2}$$

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. Furthermore, the wind speed and direction from ECMWF forecasts ~~, as well as the sea-ice concentration, thickness, drift speed and direction from TOPAZ4 forecasts.~~

~~Random forest algorithms consist of an ensemble of decision trees used for regression or classification tasks (Breiman, 2001) . The algorithms developed here are used for regression, and are based on 100 decision trees because there were no significant improvements when using more trees. The prediction from a random forest model used for regression is the mean value of the~~

results from all decision trees. For the direction of sea-ice drift, each decision tree predicts a value between 0 were calculated using equations 3 and 4, respectively, from the mean daily u and v components ($\overline{u}$ and 360. When averaging several predictions close to the northward direction , this can be an issue because values slightly higher than 0and slightly lower than 360can be averaged, possibly leading toa mean value close to $\overline{v}$ in equations 3 and 4). The computed wind direction is not the classic

280   meteorological wind direction (direction from which the wind is blowing), but the opposite (direction the wind is blowing to) in order to be consistent with the direction of sea ice drift calculated using equation 2.

$$WS = \sqrt{\overline{u}^2 + \overline{v}^2} \tag{3}$$

$$WD = \arctan 2(\overline{u}, \overline{v})\left(\frac{180}{\pi}\right) \tag{4}$$

We also tested random forest models predicting the sea ice drift along the x and y axes of the southward direction . In order

285   to avoid this issue, the predictions from all decision trees were converted to complex numbers before averaging. Then, the average of complex numbers was converted into an angle in degrees. Furthermore, the calibrated forecasts have been produced where the sea-ice concentration predicted in 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

290   was larger than 10 %.

Sensitivity tests were performed in order to optimize some parameters of the random forest algorithms. The best results (lowest mean absolute errors) were obtained by maximizing the depth of the decision trees, and by setting the number of predictor variables considered for splitting the nodes at three. These parameters have been chosen for all the models developed(which are provided by TOPAZ4 outputs) instead of the sea ice drift speed and direction. The direction and speed

295   of sea ice drift were then calculated using the start and end locations 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).

300   ### 3.3 Evaluation of sea-ice the sea ice drift directionforecasts

While comparing the speed of sea-ice sea ice drift in two data sets is straightforward, caution is needed when comparing the direction of sea-ice sea ice drift in different data sets due to the circular nature of directional data. The direction errors were calculated using equation 5, where D_x and D_y are the two directions compared in degrees (between $0°$and $360°$). Furthermore, the correlation between different data sets has been assessed using the Pearson correlation coefficient for the speed, and using

305   the circular correlation coefficient (equation 6) introduced by Fisher and Lee (1983) for the direction. In equation 6, $\bar{x}$ and $\bar{y}$

are the means of the variables x and y respectively. Similarly to the Pearson correlation coefficient, the value of the circular correlation coefficient varies between -1 and 1 (a null value indicating no correlation, 1 meaning a perfect correlation, and -1 showing a perfect anti-correlation).

$$\Delta D = D_x - D_y \Rightarrow Error = \begin{cases} \Delta D - 360, & \text{if } \Delta D > 180 \\ \Delta D + 360, & \text{if } \Delta D < -180 \\ \Delta D, & \text{otherwise} \end{cases} \tag{5}$$

$$R_c = \frac{\sum\limits_{i=1}^{n} \sin(x_i - \bar{x}) \sin(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n} \sin^2(x_i - \bar{x}) \sum\limits_{i=1}^{n} \sin^2(y_i - \bar{y})}} \tag{6}$$

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.

3.4 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).

In addition, the random forest models using all predictor variables have been compared to models in which one of the predictors was removed. This experiment was performed using the same method as the one used for determining the parameters of the random forest algorithms, which consists of training the models with data from about 80 % of the forecast start dates from the training periods, and using the remaining data for evaluating the forecasts (see section 3.1). By comparing the mean absolute errors of the different models, this method allows to determine if the predictor variables tend to improve or deteriorate the forecasts. Furthermore, it is worth noting that this method tends to underestimate the importance of highly correlated predictors since similar information is provided to the algorithm when one of the correlated predictors is removed.

4 Results

4.1 Evaluation of daily SAR observations

330 ~~The CMEMS SAR MOSAIC product provides daily averages of drift vectors derived from pairs of SAR observations. Nevertheless, the averaging process introduces uncertainties in the product because the observations are not necessarily acquired around midnight. In figure 2, this~~ The CMEMS SAR MOSAIC product has been compared to buoy observations during the period from January 2018 to ~~November 2020.~~ December 2020 in figure 2. The Pearson correlation coefficient is ~~0.78~~ 0.80 for the speed, and the circular correlation coefficient is 0.84 for the direction. The SAR observations have relatively low biases (the

335 mean error is ~~3.0~~ 2.8 degrees for the direction and ~~170 meters~~ -0.14 km / day for the speed). Furthermore, the mean absolute error is ~~21.6~~ 22.0 degrees for the direction and ~~2171 meters~~ 2.1 km / day for the speed (the mean speeds are ~~7624 and 7795 meters~~ 7.46 and 7.60 km / day for the SAR and buoy observations, respectively). The root mean square error is ~~36.0~~ 36.8 degrees for the drift direction and ~~3677 meters~~ 3.6 km / day for the drift speed. While these errors are considerable, the large number of SAR observations compared to buoy observations makes this product potentially suitable for machine learning ap-

340 plications. However, we consider these errors too large for evaluating the performances of the ~~sea-ice~~ sea ice drift forecasts using only these observations. Therefore, the performances of ~~sea-ice~~ sea ice drift forecasts have been evaluated using buoy observations, and the SAR observations have been used to study the spatial variability of the forecast performances.

4.2 Evaluation of the ~~random forest algorithms~~ calibrated forecasts

The performances of the calibrated forecasts have been evaluated and compared to those from the TOPAZ4 ~~forecasts~~ prediction

345 system during the period from June 2020 to ~~November 2020~~ May 2021 using buoy observations (figures 3~~and ??). The two versions of the calibrated forecasts outperform the TOPAZ4 forecasts for both the direction and speed of sea-ice drift, and for all lead times. ~~). For predicting the direction of ~~sea-ice~~ sea ice drift, the ~~random forest models trained using buoy observations have the best performances~~ 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

350 by these models have a mean absolute error about ~~8.1~~ 8.0 % lower than ~~the~~ TOPAZ4 forecasts~~(between 1.0 % and 16.6 % depending on the lead time). This is significantly better than the~~. The models trained with SAR observations ~~which have a mean absolute error about 5.9 % lower on average than the~~ 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

355 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 (~~57.0~~ 55.7 %) than for the models trained with SAR observations (~~54.8~~ 52.9 %). Furthermore, the ~~root mean square error is reduced by 5.4 % and 4.0 % by the calibration for the models trained with buoy and SAR observations respectively~~ 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.

360 For the speed of ~~sea-ice~~ sea ice drift, the ~~mean absolute errors are lower in the~~ models trained with buoy observations ~~during the first five days, relatively similar in both versions of the calibrated forecasts for lead times between six and seven days, and lower in~~ 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
365 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 ~~9.6 % and 7.1~~ 7.1 % and 2.5 % by the calibration for the models
trained with buoy and SAR observations, respectively. The fraction of ~~forecast~~ forecasts improved is, on average, slightly
larger for the models trained with ~~SAR observations (55.3~~ buoy observations (53.4 %) than for the models trained with ~~buoy~~
~~observations (54.9~~ SAR observations (53.1 %). ~~However, the root mean square errors are smaller in the models trained with~~
370 ~~buoy observations for all lead times. On average, the root mean square error is reduced by 12.0 % and 6.1 % by the calibration~~
~~for the models trained with buoy and SAR observations, respectively.~~ Moreover, the correlation between the buoy observations
and the ~~calibrated forecasts is higher for the models trained with buoy observations for all lead times, except for 1 and 4~~
~~days~~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
375 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 ~~November 2020~~ 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). ~~While the two versions~~
~~of the calibrated forecasts~~ Overall, both calibration methods perform relatively well for predicting the direction of ~~sea-ice~~
~~drift in most of the central Arctic , they have lower performances than TOPAZ4 in the Canadian Archipelago~~ sea ice drift in
380 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 ~~seem to perform slightly~~ 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~~ sea ice drift, the models trained with SAR observations perform better than the models trained
385 with buoy observations in most of the area analyzed. The models trained with buoy observations have particularly poor per-
formances compared to TOPAZ4 ~~in the Canadian Archipelago, along the east coast of Greenland , and north of Svalbard~~ near
the Greenland and Russian coastlines (figure 6)~~. The~~ , while the models trained with SAR observations perform better in these
areas ~~, though they also have poor performances in the Canadian Archipelago~~ (figure 7). ~~Most~~ It is worth noting that most of
the buoys taken into account for evaluating the forecasts in ~~figures 3 and ??~~ figure 3 are not located in the areas where the
390 models trained with buoy observations have poor performances, ~~and this~~ which likely explains the better performances of the
models trained with buoy observations compared to the models trained with SAR observations in ~~figures 3and ??. Furthermore,~~
~~in the Canadian Archipelago, sea-ice drift is more constrained by the coastlines than in most of the Arctic, and there is a lot of~~
~~landfast ice (Li et al., 2019). This could explain the poor performances of the random forest models that are mainly trained in~~
~~areas where the coastlines have less influence on sea-ice drift (data from the Canadian Archipelago representabout 5 % and 3~~
395 figure 3.

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 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. The sum of the relative importances of these two variables represent, on average, about 46 and 41 % of the ~~training data~~ sum of all relative importances for the models trained with ~~SAR and buoy~~ buoy and SAR observations, respectively~~).~~

4.4 ~~Importance of predictor variables~~

. However, the relative importances of these two variables decrease with increasing lead times.

~~In order to assess the importance of each predictor variable, the algorithms using all predictor variables have been compared to algorithms in which one of the predictor variables was removed (figure ??). This experiment was performed by using only data from the training periods (June 2013 - May 2020 and January 2018 - May 2020 for the algorithms~~ 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). 80 % of the data sets during these periods were selected for training the algorithms, and the remaining 20 % of the data sets were used to evaluate the performances of the algorithms. The selection of~~, respectively (figure 8). For the models predicting the speed of sea ice drift, the ~~data sets used for training and evaluating 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 from TOPAZ4 forecasts and from the observations during the initialization of the forecasts are correlated, removing one of these variables decreases the accuracy of most 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, and has been repeated 100 times in order to obtain reliable results~~. Therefore, we decided to keep both variables, even if the importances of these variables are probably underestimated due to this correlation. Furthermore, ~~the algorithms were evaluated using the same product as the one used for training the random forest models (CMEMS SAR MOSAIC product for those trained with SAR observations, and IABP buoys for those trained with buoy observations). It is worth noting that this method tends to underestimate the importance of highly correlated predictors since similar information is provided to the algorithms when one of the correlated predictors is removed~~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 ~~both calibration methods, the most important variable for predicting the drift direction is the sea-ice~~ the models predicting the direction of sea ice drift, removing the drift direction from TOPAZ4 forecasts ~~. Similarly, the sea-ice drift speed from~~

~~TOPAZ4 is the most important variable for~~ 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. Moreover, the ECMWF wind forecasts and~~ 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 ~~geographical coordinates in the stereographic projection used in the TOPAZ4 system also have a significant impact on the forecast errors for predicting both the direction and speed of sea-ice drift~~ speed of sea ice drift for a lead time of 4 days trained with SAR observations. For the ~~algorithms~~ other models predicting the speed of ~~sea-ice drift,~~ 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 ~~are also significantly~~ for the speed of sea ice drift are also considerably reduced by adding the ~~sea-ice~~ sea ice thickness forecasts from TOPAZ4 ~~. Some predictor variables do not improve the forecast accuracy for all lead times. In particular, for the algorithms predicting the speed of sea-ice drift trained with buoy observations, the forecasts of wind direction and drift direction increase the mean absolute errors for five and four lead times, respectively~~(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).

5 Discussion and conclusion

The characteristics and performances of the calibrated forecasts developed in this study depend on the observations used for training the algorithms, as well as the data sets used for the predictor variables. The 24-hour mean composites of drift vectors provided by the CMEMS SAR MOSAIC product have been used for training some algorithms due to the fixed time spans of this data set, the large number of available observations, and the relatively high ~~spatiotemporal~~ spatio-temporal resolution compared to ~~sea-ice~~ sea ice drift products developed from passive microwave observations (Lavergne et al., 2010; Girard-Ardhuin and Ezraty, 2012; Tschudi et al., 2020). However, the spatial resolution (10 km) and the time spans (24 hours) of this product prevent to develop high-resolution calibrated forecasts. Furthermore, the averaging process of drift vectors introduces significant uncertainties in the product (figure 2), which represent a limitation for developing accurate calibrated forecasts. ~~On the other hand, buoy~~ Buoy observations are more accurate, but the relatively low number of available observations is a limitation for the development of random forest ~~algorithms~~models. Both methods also have common limitations such as the heterogeneous spatio-temporal sampling of the buoy and SAR observations. The random forest algorithms are more influenced by areas often covered by ~~sea-ice~~ sea ice drift observations than by areas with a poor coverage in these observations. This could potentially explain some of the spatial variability of the performances of the calibrated forecasts. ~~In particular, the poor performances of the calibrated forecasts compared to TOPAZ4 in the Canadian Archipelago could be partly due to local characteristics (narrow channels and presence of landfast ice) that are not often encountered in~~ Overall, the calibrated forecasts have their best performances in the Central Arctic where most of the training data ~~sets~~are located.

The forecasts used for the predictor variables are produced by operational models (ECMWF Integrated Forecasting System
and the TOPAZ4 prediction system) which are constantly developed. The development of these systems could affect the performances of the random forest algorithms due to changes between the different versions of the models. This issue is more important for the ~~algorithms~~ models trained with buoy observations due to the longer period used for training these ~~algorithms (June 2013 to May 2020~~ models. 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, and using a shorter training period would have improved the performances for predicting the speed of sea ice drift (figure S2 of the supplementary material). Furthermore, while only the TOPAZ4 forecasts starting on Thursdays were used for training the random forest algorithms (only these forecasts are stored in the long-term archive), the operational forecasts have been produced daily. Because ~~only the forecasts starting on Thursdays are initialized using data assimilation~~ 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).

Despite these limitations, it has been shown that the ~~calibrated forecasts produced by both methods outperform the~~ random forest models trained with buoy observations outperform TOPAZ4 forecasts for both the speed and direction of sea ice drift, except for 10-day lead time for the direction of sea ice drift. The models trained with SAR observations significantly outperform TOPAZ4 forecasts for ~~both the direction and speed of sea-ice drift when using buoy observations as reference (figures 3 and ??)~~ lead times up to 5 days for the direction of sea ice drift, but have similar performances as TOPAZ4 forecasts for longer lead times. For the speed of sea ice drift, the models trained with SAR observations have slightly larger mean absolute errors than TOPAZ4 for lead times up to 5 days, but significantly outperform TOPAZ4 forecasts for longer lead times. On average, the mean absolute errors for the direction of ~~sea-ice drift are 8.1 % and 5.9~~ sea ice drift are 8.0 % and 3.3 % lower than in TOPAZ4 forecasts for the ~~algorithms~~ models trained with buoy and SAR observations, respectively. For the models predicting the speed of ~~sea-ice~~ sea ice drift, the mean absolute errors are reduced by ~~9.6 % and 7.1~~ 7.1 % and 2.5 % for the ~~algorithms~~ models trained with buoy and SAR observations, respectively. The lower errors of the ~~algorithms~~ models trained with buoy observations, despite their smaller training data sets, show that the accuracy of the observations used for the target variables plays a crucial role in the performances of the calibrated forecasts.

The spatial analysis of the forecast performances using SAR observations as reference has shown that the ~~algorithms~~ models trained with buoy observations ~~slightly outperform the algorithms~~ outperform the models trained with SAR observations for predicting the direction of ~~sea-ice~~ sea ice drift in most of the area analyzed in this study (figures 4 and 5). However, for the algorithms predicting the speed of ~~sea-ice~~ sea ice drift, the ~~forecasts~~ models trained with SAR observations have better performances in most of the area analyzed in this study (figures 6 and 7). The ~~algorithms~~ models trained with buoy observations have particularly poor performances ~~north of Svalbard, along the east coast of Greenland, and in the Canadian Archipelago~~ for predicting the speed of sea ice drift near the Greenland and Russian coastlines. This is likely due to the low number of buoy observations available for training the algorithms in these areas and the particular characteristics of these areas (~~presence of~~

~~landfast ice in the Canadian Archipelago, and high sea-ice~~ high drift speed along the ~~east coast of Greenland and north of Svalbard~~ Greenland East coast and presence of landfast ice along the Russian coast).

500   In order to reduce forecast errors from numerical prediction systems, calibration procedures can be applied, although their performances depend on the number and accuracy of the observations used as target variables. The increasing amount of satellite data available and the improvements in ~~sea-ice~~ sea ice remote sensing, as well as the development of new statistical approaches, enhance the potential of calibration techniques for ~~sea-ice~~ sea ice forecasting. This should contribute to improve the accuracy of ~~sea-ice~~ sea ice forecasts delivered to maritime end-users.

505   *Code and data availability.* Buoy observations are available on the International Arctic Buoy Programme (IABP) website (https://iabp.apl.uw.edu). Synthetic-aperture radar observations and TOPAZ4 forecasts are available on the Copernicus Marine Environment Monitoring Service server (https://resources.marine.copernicus.eu/). The OSI-SAF sea ice concentration observations can be downloaded from the MET-Norway FTP server (ftp://osisaf.met.no/ reprocessed/ice/conc/v2p0) until 2015, but the data after 2015 are not publicly available. A licence is needed to download the wind forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). Furthermore, the codes used for

510   this analysis are available in the following Github directory: "https://github.com/cyrilpalerme/Calibration_of_sea_ice_drift_forecasts/".

Author contributions. C.P. conducted the analysis, and wrote the majority of the manuscript. C.P. and M.M. designed the research and contributed to the discussions of the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank John Bjørnar Bremnes, Siri Sofie Eide, Roberto Saldo, and Thomas Lavergne for

515   valuable discussions, as well as Atle Macdonald Sørensen for maintaining the production of ~~sea-ice~~ sea ice concentration observations from AMSR2 and SSMIS at the Norwegian Meteorological Institute. The authors received funding from the Copernicus Marine Environmental and Monitoring Service Arctic Marine Forecasting Center and IcySea project through Mercator Océan, as well as, from the SALIENSEAS project funded by the Norwegian Research Council contract number 276223. This is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological

520   Organisation (WMO). Finally, we thank the three reviewers for their comments which significantly improved the manuscript.

References

Ahijevych, D., Pinto, J. O., Williams, J. K., and Steiner, M.: Probabilistic Forecasts of Mesoscale Convective System Initiation Using the Random Forest Data Mining Technique, Weather and Forecasting, 31, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1, 2016.

Berkman, P. A., Fiske, G., Røyset, J.-A., Brigham, L. W., and Lorenzini, D.: Next-Generation Arctic Marine Shipping Assessments, pp. 241–268, Springer International Publishing, https://doi.org/10.1007/978-3-030-25674-6_11, 2020.

Bleck, R.: An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates, Ocean Modelling, 4, 55–88, https://doi.org/https://doi.org/10.1016/S1463-5003(01)00012-9, 2002.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Cavalieri, D. J. and Parkinson, C. L.: Arctic sea ice variability and trends, 1979-2010, The Cryosphere, 6, 881–889, https://doi.org/10.5194/tc-6-881-2012, 2012.

Chassignet, E. P., Hurlburt, H. E., Smedstad, O. M., Halliwell, G. R., Hogan, P. J., Wallcraft, A. J., and Bleck, R.: Ocean Prediction with the Hybrid Coordinate Ocean Model (HYCOM), pp. 413–426, Springer Netherlands, Dordrecht, https://doi.org/10.1007/1-4020-4028-8_16, 2006.

Chi, J. and Kim, H.-c.: Prediction of arctic sea ice concentration using a fully data driven deep neural network, Remote Sensing, 9, 1305, 2017.

Comeau, D., Giannakis, D., Zhao, Z., and Majda, A. J.: Predicting regional and pan-Arctic sea ice anomalies with kernel analog forecasting, Climate Dynamics, 52, 5507–5525, 2019.

Comiso, J. C., Meier, W. N., and Gersten, R.: Variability and trends in the Arctic Sea ice cover: Results from different techniques, Journal of Geophysical Research: Oceans, 122, 6883–6900, https://doi.org/https://doi.org/10.1002/2017JC012768, 2017.

Eriksen, T. and Olsen, Ø.: Vessel Tracking Using Automatic Identification System Data in the Arctic, pp. 115–136, Springer International Publishing, https://doi.org/10.1007/978-3-319-78425-0_7, 2018.

Fisher, N. I. and Lee, A. J.: A correlation coefficient for circular data, Biometrika, 70, 327–332, https://doi.org/10.1093/biomet/70.2.327, 1983.

Fritzner, S., Graversen, R., and Christensen, K. H.: Assessment of High-Resolution Dynamical and Machine Learning Models for Prediction of Sea Ice Concentration in a Regional Application, Journal of Geophysical Research: Oceans, 125, e2020JC016 277, https://doi.org/https://doi.org/10.1029/2020JC016277, 2020.

Gagne, David John, I., 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.

Gegiuc, A., Similä, M., Karvonen, J., Lensu, M., Mäkynen, M., and Vainio, J.: Estimation of degree of sea ice ridging based on dual-polarized C-band SAR data, The Cryosphere, 12, 343–364, https://doi.org/10.5194/tc-12-343-2018, 2018.

Girard-Ardhuin, F. and Ezraty, R.: Enhanced Arctic Sea Ice Drift Estimation Merging Radiometer and Scatterometer Data, IEEE Transactions on Geoscience and Remote Sensing, 50, 2639–2648, https://doi.org/10.1109/TGRS.2012.2184124, 2012.

Han, H., Im, J., Kim, M., Sim, S., Kim, J., Kim, D.-j., and Kang, S.-H.: Retrieval of Melt Ponds on Arctic Multiyear Sea Ice in Summer from TerraSAR-X Dual-Polarization Data Using Machine Learning Approaches: A Case Study in the Chukchi Sea with Mid-Incidence Angle Data, Remote Sensing, 8, 57, https://doi.org/10.3390/rs8010057, 2016.

Hebert, D. A., Allard, R. A., Metzger, E. J., Posey, P. G., Preller, R. H., Wallcraft, A. J., Phelps, M. W., and Smedstad, O. M.: Short-term sea ice forecasting: An assessment of ice concentration and ice drift forecasts using the U.S. Navy's Arctic Cap Nowcast/Forecast System, Journal of Geophysical Research: Oceans, 120, 8327–8345, https://doi.org/10.1002/2015JC011283, 2015.

560 Herman, G. R. and Schumacher, R. S.: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests, Monthly Weather Review, 146, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1, 2018.

Hunke, E. C. and Dukowicz, J. K.: An Elastic–Viscous–Plastic Model for Sea Ice Dynamics, Journal of Physical Oceanography, 27, 1849 – 1867, https://doi.org/10.1175/1520-0485(1997)027<1849:AEVPMF>2.0.CO;2, 1997.

Jeuring, J. and Knol-Kauffman, M.: Mapping Weather, Water, Ice and Climate Knowledge & Information Needs for Maritime Activities in the Arctic. Survey Report, 2019.

565 Kim, Y. J., Kim, H.-C., Han, D., Lee, S., and Im, J.: Prediction of monthly Arctic sea ice concentrations using satellite and reanalysis data based on convolutional neural networks, The Cryosphere, 14, 1083–1104, https://doi.org/10.5194/tc-14-1083-2020, https://tc.copernicus.org/articles/14/1083/2020/, 2020.

Kwok, R.: Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018), Environmental Research Letters, 13, 105 005, https://doi.org/10.1088/1748-9326/aae3ec, https://doi.org/10.1088/1748-9326/aae3ec, 2018.

570 Kwok, R. and Rothrock, D. A.: Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008, Geophysical Research Letters, 36, https://doi.org/https://doi.org/10.1029/2009GL039035, 2009.

Lavergne, T., Eastwood, S., Teffah, Z., Schyberg, H., and Breivik, L.-A.: Sea ice motion from low-resolution satellite sensors: An alternative method and its validation in the Arctic, Journal of Geophysical Research: Oceans, 115, https://doi.org/https://doi.org/10.1029/2009JC005958, 2010.

575 Lavergne, T., Sørensen, A. M., Kern, S., Tonboe, R., Notz, D., Aaboe, S., Bell, L., Dybkjær, G., Eastwood, S., Gabarro, C., Heygster, G., Killie, M. A., Brandt Kreiner, M., Lavelle, J., Saldo, R., Sandven, S., and Pedersen, L. T.: Version 2 of the EUMETSAT OSI SAF and ESA CCI sea-ice concentration climate data records, The Cryosphere, 13, 49–78, https://doi.org/10.5194/tc-13-49-2019, 2019.

Lee, S., Im, J., Kim, J., Kim, M., Shin, M., Kim, H.-c., and Quackenbush, L.: Arctic Sea Ice Thickness Estimation from CryoSat-2 Satellite Data Using Machine Learning-Based Lead Detection, Remote Sensing, 8, 698, https://doi.org/10.3390/rs8090698, 2016.

580 Li, Z., Zhao, J., Su, J., Li, C., Cheng, B., Hui, F., Yang, Q., and Shi, L.: Spatial and Temporal Variations in the Extent and Thickness of Arctic Landfast Ice, Remote Sensing, 12, 64, https://doi.org/10.3390/rs12010064, 2019.

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.

Mao, Y. and Sorteberg, A.: Improving radar based precipitation nowcasts with machine learning using an approach based on random forest, 585 Weather and Forecasting, pp. 1–51, https://doi.org/10.1175/WAF-D-20-0080.1, 2020.

Miao, X., Xie, H., Ackley, S. F., Perovich, D. K., and Ke, C.: Object-based detection of Arctic sea ice and melt ponds using high spatial resolution aerial photographs, Cold Regions Science and Technology, 119, 211–222, https://doi.org/https://doi.org/10.1016/j.coldregions.2015.06.014, 2015.

Olason, E. and Notz, D.: Drivers of variability in Arctic sea-ice drift speed, Journal of Geophysical Research: Oceans, 119, 5755–5775, 590 https://doi.org/10.1002/2014JC009897, 2014.

Park, J.-W., Korosov, A. A., Babiker, M., Won, J.-S., Hansen, M. W., and Kim, H.-C.: Classification of sea ice types in Sentinel-1 synthetic aperture radar images, The Cryosphere, 14, 2629–2645, https://doi.org/10.5194/tc-14-2629-2020, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830, 2011.

595 Petty, A. A., Kurtz, N. T., Kwok, R., Markus, T., and Neumann, T. A.: Winter Arctic Sea Ice Thickness From ICESat-2 Freeboards, Journal of Geophysical Research: Oceans, 125, e2019JC015 764, https://doi.org/https://doi.org/10.1029/2019JC015764, 2020.

Rabatel, M., Rampal, P., Carrassi, A., Bertino, L., and Jones, C. K. R. T.: Impact of rheology on probabilistic forecasts of sea ice trajectories: application for search and rescue operations in the Arctic, The Cryosphere, 12, 935–953, https://doi.org/10.5194/tc-12-935-2018, 2018.

Rampal, P., Weiss, J., and Marsan, D.: Positive trend in the mean speed and deformation rate of Arctic sea ice, 1979–2007, Journal of
600 Geophysical Research: Oceans, 114, https://doi.org/10.1029/2008JC005066, 2009.

Ricker, R., Hendricks, S., Kaleschke, L., Tian-Kunze, X., King, J., and Haas, C.: A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data, The Cryosphere, 11, 1607–1623, https://doi.org/10.5194/tc-11-1607-2017, https://tc.copernicus.org/articles/11/1607/2017/, 2017.

Sakov, P., Counillon, F., Bertino, L., Lisæter, K. A., Oke, P. R., and Korablev, A.: TOPAZ4: an ocean-sea ice data assimilation system for the
605 North Atlantic and Arctic, Ocean Science, 8, 633–656, https://doi.org/10.5194/os-8-633-2012, 2012.

Schweiger, A. J. and Zhang, J.: Accuracy of short-term sea ice drift forecasts using a coupled ice-ocean model, Journal of Geophysical Research: Oceans, 120, 7827–7841, https://doi.org/10.1002/2015JC011273, 2015.

Spreen, G., Kwok, R., and Menemenlis, D.: Trends in Arctic sea ice drift and role of wind forcing: 1992–2009, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL048970, 2011.

610 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.

Tandon, N. F., Kushner, P. J., Docquier, D., Wettstein, J. J., and Li, C.: Reassessing Sea Ice Drift and Its Relationship to Long-Term Arctic Sea Ice Loss in Coupled Climate Models, Journal of Geophysical Research: Oceans, 123, 4338–4359, https://doi.org/10.1029/2017JC013697, 2018.

615 Tschudi, M. A., Meier, W. N., and Stewart, J. S.: An enhancement to sea ice motion and age products at the National Snow and Ice Data Center (NSIDC), The Cryosphere, 14, 1519–1536, https://doi.org/10.5194/tc-14-1519-2020, 2020.

Wang, L., Yuan, X., Ting, M., and Li, C.: Predicting Summer Arctic Sea Ice Concentration Intraseasonal Variability Using a Vector Autoregressive Model, Journal of Climate, 29, 1529–1543, https://doi.org/10.1175/JCLI-D-15-0313.1, 2016.

Wang, L., Yuan, X., and Li, C.: Subseasonal forecast of Arctic sea ice concentration via statistical approaches, Climate Dynamics, 52,
620 4953–4971, 2019.

Williams, T., Korosov, A., Rampal, P., and Ólason, E.: Presentation and evaluation of the Arctic sea ice forecasting system neXtSIM-F, The Cryosphere Discussions, 2019, 1–31, https://doi.org/10.5194/tc-2019-154, https://tc.copernicus.org/preprints/tc-2019-154/, 2019.

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 Science, 13, 123–144, https://doi.org/10.5194/os-13-123-2017, 2017.

625 Yu, X., Rinke, A., Dorn, W., Spreen, G., Lüpkes, C., Sumata, H., and Gryanik, V. M.: Evaluation of Arctic sea ice drift and its dependency on near-surface wind and sea ice conditions in the coupled regional climate model HIRHAM–NAOSIM, The Cryosphere, 14, 1727–1746, https://doi.org/10.5194/tc-14-1727-2020, 2020.
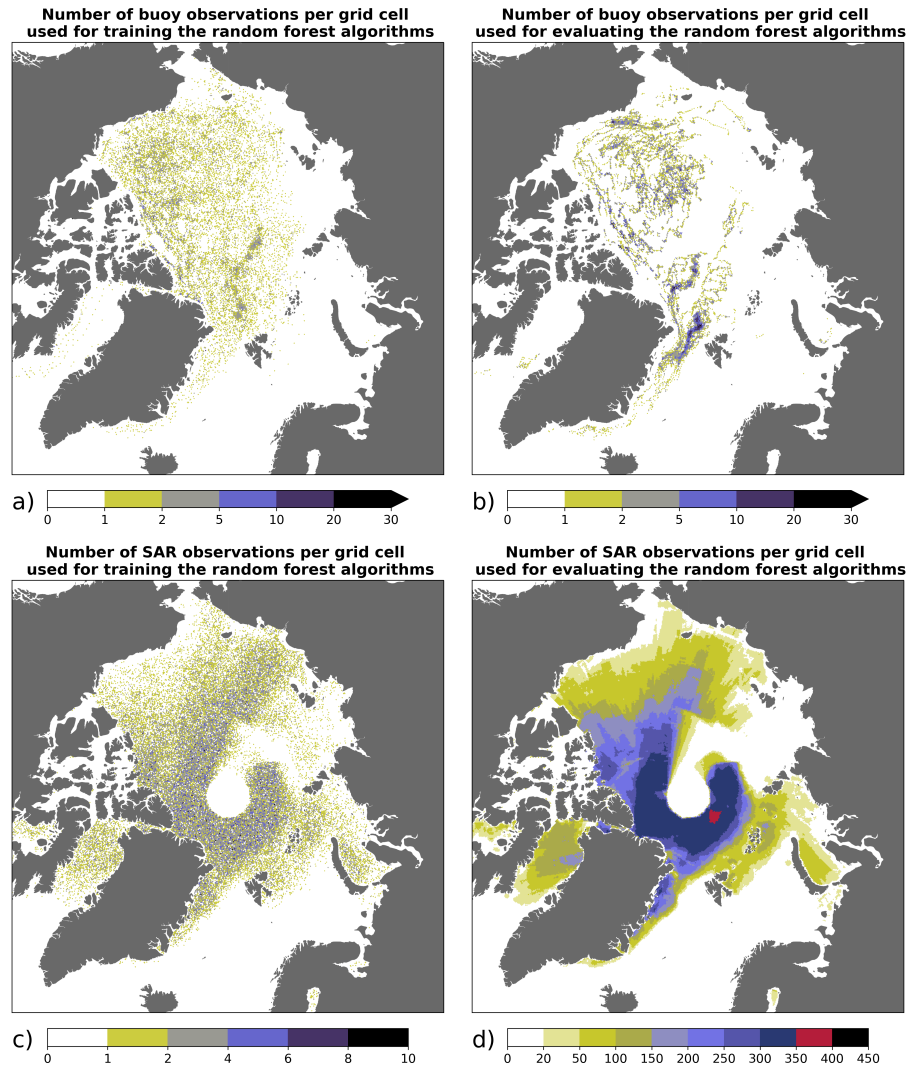
Figure 1. a) Number of buoy observations per grid cell used for training the random forest algorithms during the period from June 2013 to May 2020. b) Number of buoy observations per grid cell used for evaluating the random forest algorithms during the period from June 2020 to ~~November 2020.~~ May 2021. c) Number of SAR observations per grid cell used for training the random forest algorithms during the period from January 2018 to May 2020. d) Number of SAR observations per grid cell used for evaluating the random forest algorithms during the period from June 2020 to ~~November 2020.~~ May 2021. These four maps show the number of observations for a lead time of one day (similar results have been found for other lead times).
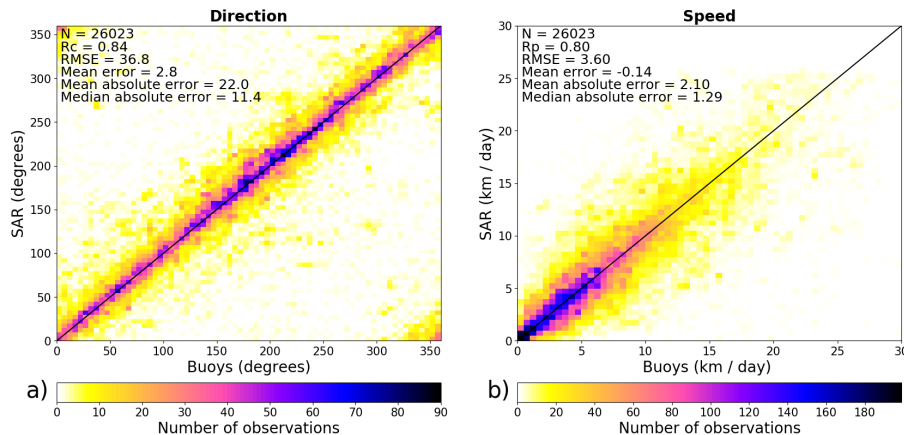
Figure 2. ~~Difference in mean absolute errors during the period from June 2020 to November 2020 between the algorithms trained using all the grid points with SAR observations and the algorithms trained using only a fraction~~ Evaluation of the ~~grid points with SAR observations~~ CMEMS SEAICE_GLO_SEAICE_L4_NRT_OBSERVATIONS_011_006 product (MOSAIC version 2.0) for the sea ice drift direction (a) and ~~the~~ speed (b) ~~of sea-ice drift. IAPB buoy observations have been used as reference. Positive values mean that~~ during the ~~algorithms using only a fraction of~~ period from January 2018 to December 2020. The axes are bounded to 30 km / day for the ~~available grid points~~ speed for ~~training outperform~~ clarity, but the ~~algorithms using all~~ speed of 73 buoys exceed 30 km / day. However, these buoys were taken into account when calculating the ~~available grid points~~ statistics shown on the figures (N: number of samples, Rc: circular correlation coefficient, Rp: Pearson correlation coefficient, RSME: root mean square error). The ~~legend from figure a) shows~~ color scales represent the ~~fraction~~ number of ~~grid points randomly selected~~ observations in each bin of 5 degrees for ~~training~~ the ~~random forest algorithms~~ direction and 0.5 km / day for the speed.

~~Evaluation of the CMEMS SEAICE_GLO_SEAICE_L4_NRT_OBSERVATIONS_011_006 product (MOSAIC version 2.0) for the sea-ice drift direction (a) and speed (b) during the period from January 2018 to November 2020. The axes are bounded to 30 km / day for the speed for clarity, but the speed of 72 buoys exceed 30 km / day. However, these buoys were taken into account when calculating the statistics shown on the figures (N: number of samples, Rc: circular correlation coefficient, Rp: Pearson correlation coefficient, RSME: root mean square error). The color scales represent the number of observations in each bin of 5 degrees for the direction and 500 meters / day for the speed.~~
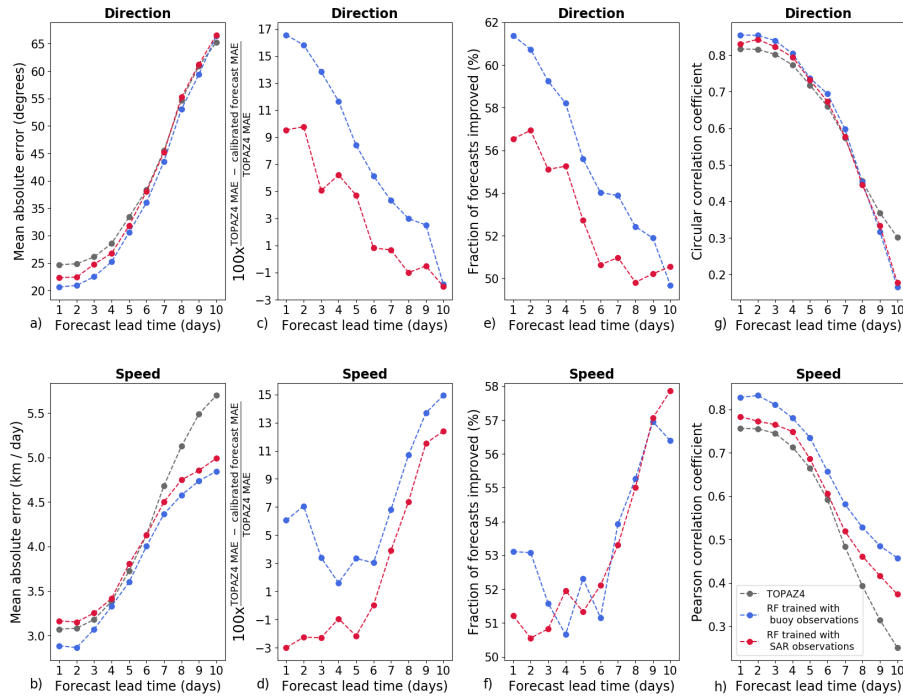
Figure 3. Evaluation of the performances of the calibrated forecasts produced using random forest (RF) algorithms and the TOPAZ4 forecasts during the period from June 2020 to ~~November 2020. IAPB~~ May 2021. IABP buoy observations have been used as reference. Mean absolute errors (MAE) of the forecasts for the direction (a) and the speed (b), ~~fraction~~ and relative improvement ($100 \times \frac{\text{TOPAZ4 MAE - calibrated forecast MAE}}{\text{TOPAZ4 MAE}}$) for the direction (c) and the speed (d). Fraction of calibrated forecasts with lower absolute errors than the TOPAZ4 forecasts for the direction (~~c~~e) and the speed (~~d~~f). ~~Relative improvement ($100 \times \frac{\text{TOPAZ4 errors - calibrated forecast errors}}{\text{TOPAZ4 errors}}$)~~ Circular and Pearson correlation coefficients between the forecasts and the buoy observations for the direction (~~e~~g) and the speed (f), respectively.

Calibrated forecasts (models trained with buoy observations)
Fraction of forecasts improved for the direction of sea-ice drift (%)
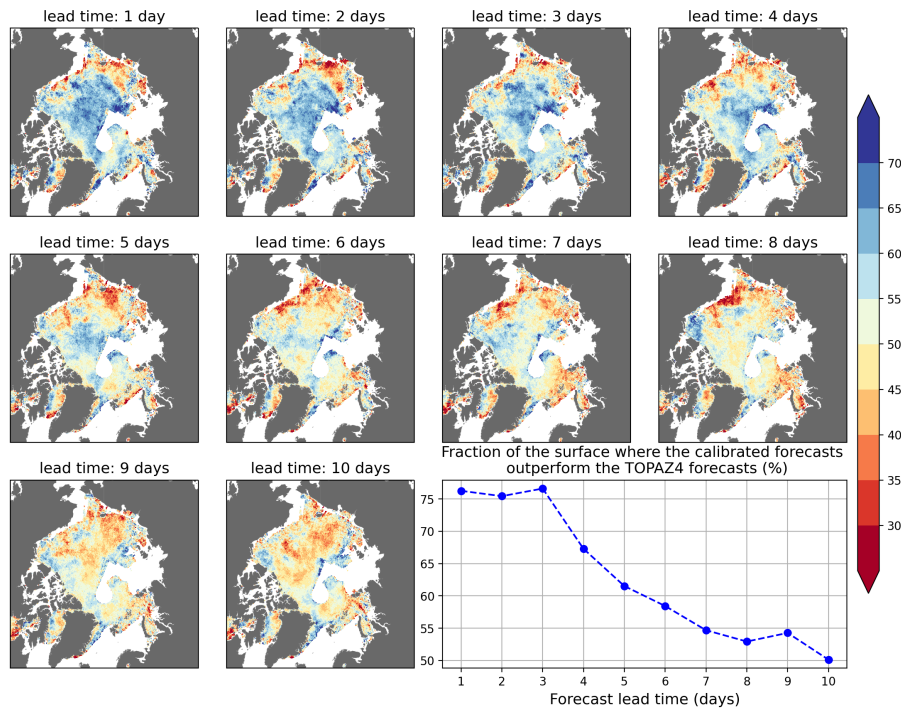
Figure 4. ~~Evaluation of the performances~~ Fraction of ~~the~~ calibrated forecasts ~~and~~ trained using buoy observations which outperform the TOPAZ4 forecasts for the direction of sea ice drift during the period from June 2020 to ~~November 2020. IAPB buoy~~ May 2021. Daily SAR observations have been used as reference. ~~Root mean square errors of~~ The graph in the ~~forecasts for~~ lower right corner shows the ~~direction (a) and for~~ fraction of the ~~speed (b), circular correlation coefficient for~~ surface where the ~~direction (c) and Pearson correlation coefficient for~~ fraction of calibrated forecasts outperforming the ~~speed (d)~~ TOPAZ4 forecasts is higher or equal to 50 %.

~~Fraction of calibrated forecasts trained using buoy observations which outperform the TOPAZ4 forecasts for the direction of sea-ice drift during the period from June 2020 to November 2020. Daily SAR observations have been used as reference. The graph in the lower right corner shows the fraction of the surface where the fraction of calibrated forecasts outperforming the TOPAZ4 forecasts is higher or equal to 50 %.~~

Figure 5. Fraction of calibrated forecasts trained using SAR observations which outperform the TOPAZ4 forecasts for the direction of ~~sea-ice~~ sea ice drift during the period from June 2020 to ~~November 2020.~~ May 2021. Daily SAR observations have been used as reference. The graph in the lower right corner shows the fraction of the surface where the fraction of calibrated forecasts outperforming the TOPAZ4 forecasts is higher or equal to 50 %.

Calibrated forecasts (models trained with buoy observations)
Fraction of forecasts improved for the speed of sea-ice drift (%)
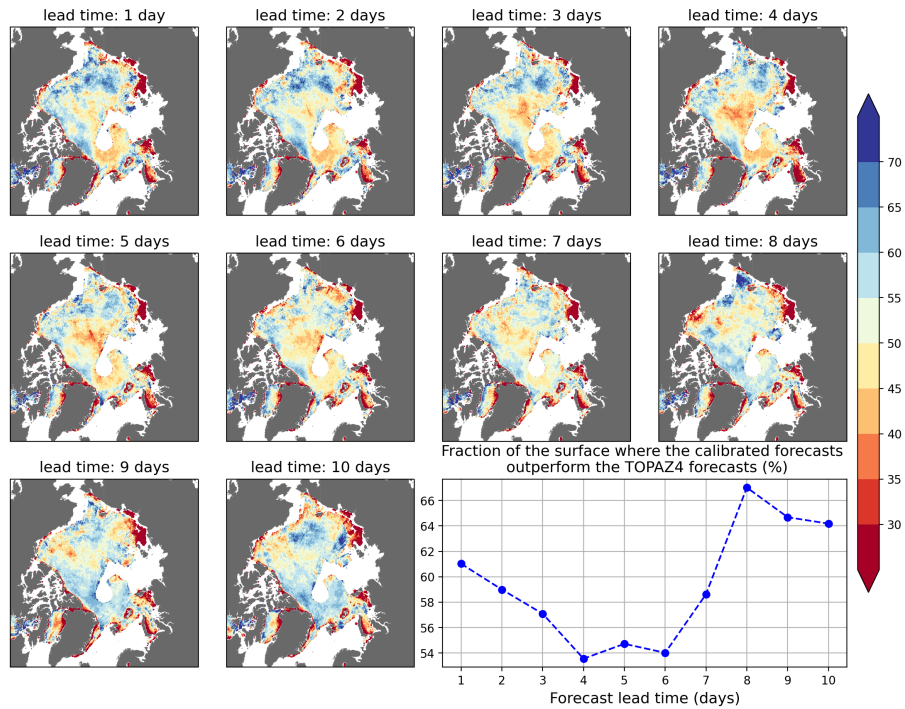
Figure 6. Fraction of calibrated forecasts trained using buoy observations which outperform the TOPAZ4 forecasts for the speed of ~~sea-ice~~ sea ice drift during the period from June 2020 to ~~November 2020.~~ May 2021. Daily SAR observations have been used as reference. The graph in the lower right corner shows the fraction of the surface where the fraction of calibrated forecasts outperforming the TOPAZ4 forecasts is higher or equal to 50 %.

Figure 7. Fraction of calibrated forecasts trained using SAR observations which outperform the TOPAZ4 forecasts for the speed of ~~sea-ice~~ sea ice drift during the period from June 2020 to ~~November 2020.~~ May 2021. Daily SAR observations have been used as reference. The graph in the lower right corner shows the fraction of the surface where the fraction of calibrated forecasts outperforming the TOPAZ4 forecasts is higher or equal to 50 %.
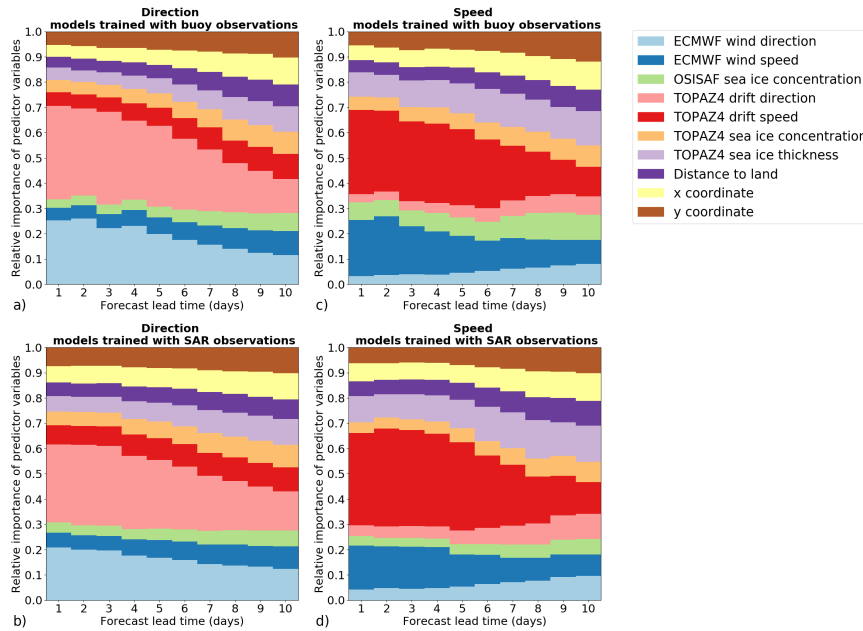
Figure 8. ~~Differences in mean absolute error when one~~ Relative importance of the predictor variables ~~is not used in the random forest algorithms~~ for the direction (a, b) and the speed (c, d) of ~~sea-ice~~ sea ice drift ~~. The results are shown for the algorithms trained with buoy observations (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~~ assessed 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~~impurity-based feature importance method.
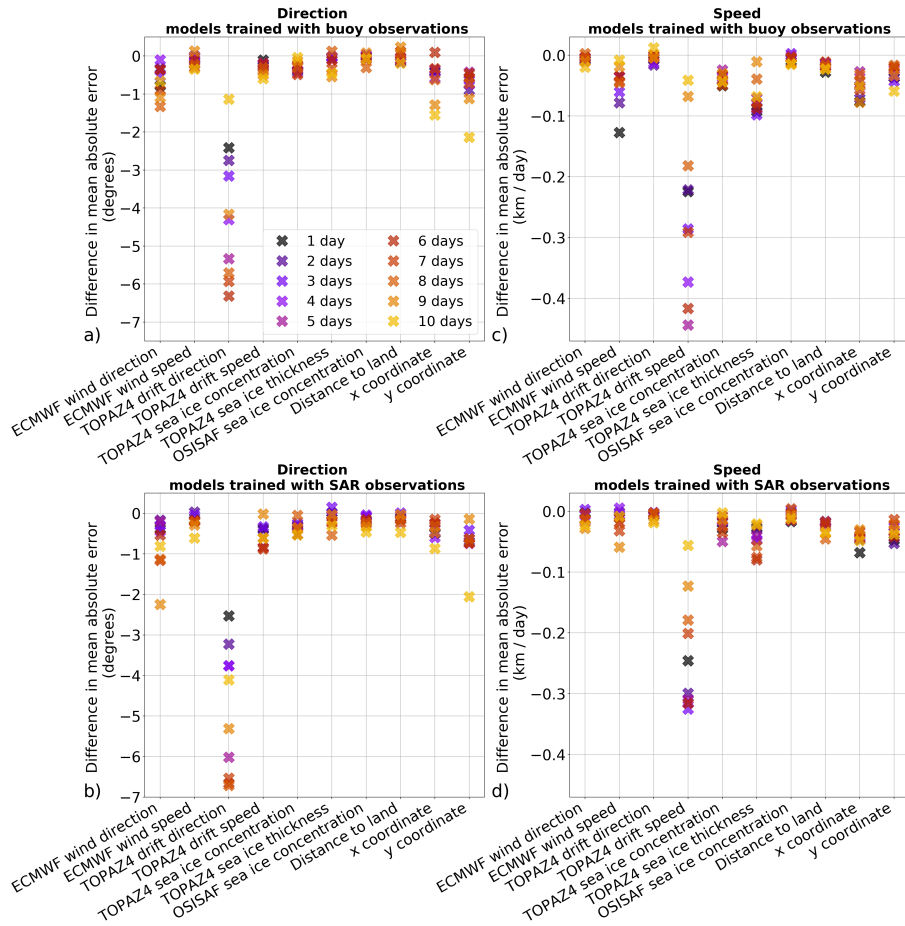
Figure 9. Differences in mean absolute error when one of the predictor variables is not used in the random forest models for the direction (a, b) and speed (c, d) of sea ice drift. The results are shown for the models trained with buoy observations (a, c), and for the models trained with SAR observations (b, d). The lead times are indicated in the legend of figure a). The differences represent the subtraction between the performances of the models using all the predictor variables and the models in which one predictor variable was not used. Therefore a negative value means that adding the variable in the algorithm improves the forecasts.