The Cryosphere Discussions

The Cryosphere Discuss., referee comment RC1
https://doi.org/10.5194/tc-2021-24-RC1, 2021

Comment on tc-2021-24

Anonymous Referee #1

Referee comment on "Calibration of sea ice drift forecasts using random forest algorithms" by Cyril Palerme and Malte Müller, The Cryosphere Discuss., https://doi.org/10.5194/tc-2021-24-RC1, 2021

##########
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? I would be good to provide some clarification and/or discussion in this regard.

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.

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.

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.

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?

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

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?

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?)

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

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?

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

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? 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? 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?

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. Furthermore, it is not surprising that the TOPAZ4 dirft 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.

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?