Replies to Reviewer 1

This manuscript describes how including sea ice deformation derived from a satellite data product of sea ice drift may improve short sea ice predictive skills of an Arctic sea ice model. The main "trick" is to connect derived deformation to scalar model variables with some "memory", i.e ice concentration and damage, as assimilating the deformation or drift directly is known to be problematic as this information contains little memory and is usually lost within time steps of the model. The data assimilation (DA) itself is a simple re-initialisation scheme for ice concentration and damage. The authors speculate how their findings can be used in more sophisticated DA systems.

The topic is interesting and important as sea ice forecasts are thought to become more and more relevant for Arctic shipping and exploration. There are some issues with the manuscript that require careful rewriting and even repeating the experiments, so that I cannot recommend publication in the present form.

We appreciate the reviewer's constructive comments. As requested by the reviewer, more experiments for evaluating predictability of sea ice deformation were run. We have also improved the clarity of the mathematical formulation of the problem. The text below, which describes these runs and defines the terminology, is added to the Methodology section. See also Fig. 1 for further explanation.

Let $\mathbf{x}$ denote the model state variables (e.g., concentration, damage, etc.). Experiments start from 1 December 2020 (t_0) and last for two months. $\mathbf{x}_{t_0}$ is the initial condition. $M_{t_n \to t_{n+1}}$ is the non-linear model (neXtSIM) to propagate state from time t_n to t_{n+1}.

Let $\mathbf{y}$ denote the observations (total deformation rate), which is related to the model state variables through $\mathbf{y}_t = H(\mathbf{x}_t)$, where H is the observation operator. Real satellite observations $\mathbf{y}_t^o$ are available throughout the test period. Although the deformation rate is derived from sea ice drift, derived from Radarsat-2 SAR images, we call them "observations of deformation" as opposed to "simulation of deformation" by neXtSIM.

In the first experiment a verifying "truth run" is generated:

$\mathbf{x}_t^{\mathrm{tr}} = M_{t_0 \to t}(\mathbf{x}_{t_0})$

The period before 1 January 2021 is used as a spin-up time and the data from $\mathbf{x}_t^{\mathrm{tr}}$ is not used.

Then four sets of 10-days forecasts are ran every day:

1. Forecasts initiated from truth:

$\mathbf{x}_{t_1 \to t}^{\mathrm{tr}} = M_{t_1 \to t}(\mathbf{x}_{t_1}^{\mathrm{tr}}) + \boldsymbol{\psi}_t,$

where $\boldsymbol{\psi}_t$ is a random noise due to uncertainties in model numerics that cause the forecasts run to differ from the truth run. These forecasts are evaluated by computing the error in observation space:

$\varepsilon_T = \left\langle H(\mathbf{x}_{t \to t+\delta t}^{\mathrm{tr}}) - \mathbf{y}_{t+\delta t}^{tr} \right\rangle,$

where $\langle \cdot \rangle$ denotes averaging over the different forecasts starting from $t_1, t_2, \ldots, t_n$, (i.e. $\langle e_{t \to t+\delta t} \rangle = \sum_{t=t_1}^{t_n} e_{t \to t+\delta t}$, then plotted w.r.t. lead time δt).

2. Forecasts without data assimilation:

$\mathbf{x}_{t_1 \to t} = M_{t_1 \to t}(\mathbf{x}_{t_1}) + \boldsymbol{\psi}_t.$

The first forecast is initiated from t_0, subsequent forecasts are initiated from the outputs of the previous forecasts. The forecasts initiated during the spin-up period are not used. The forecasts after 1 January are evaluated against truth:

$\varepsilon_t^B = H(\mathbf{x}_t) - \mathbf{y}_t^{tr}$,

and against real observations:

$\varepsilon_t^O = H(\mathbf{x}_t) - \mathbf{y}_t^o.$

During the spin-up period ε_t^B grows and reaches its saturation level ε_B, which we consider to be the climatological level for this error. Since the forecasts without data assimilation don't see real data, the error ε_t^O averaged over one month (ε_O) can also be considered as the climatological level.

3. Forecasts with assimilation of synthetic data:

$$\mathbf{x}^{as}_{t_1\to t} = M_{t_1\to t}(\mathbf{x}^{as}_{t_1}) + \boldsymbol{\psi}_t,$$

where $\mathbf{x}^{as}_{t_1}$ is the analysis of synthetic observations from the truth run and the forecasts without assimilation performed at t_1: $\mathbf{x}^{as}_{t_1} = A(\mathbf{x}_{t_1}, \mathbf{y}^{tr}_{t_1}; H', \mathbf{w})$. In the assimilation scheme in this paper, we use the inverse operator H' to compute model state (concentration and damage) from the observed deformation, $\mathbf{x}_t = H'(\mathbf{y}_t)$ (see Eqs. 9 and 10 for how H' is constructed) and $\mathbf{w}$ are the tuning parameters (see Eqs. 11 - 13 for how A is constructed). These forecasts are evaluated with:

$$\varepsilon_S = \left\langle H(\mathbf{x}^{as}_{t\to t+\delta t}) - \mathbf{y}^{tr}_{t+\delta t} \right\rangle,$$

4. Forecasts with assimilation of real satellite data:

$$\mathbf{x}^{ar}_{t_1\to t} = A(\mathbf{x}_{t_1}, \mathbf{y}^{o}_{t_1}; H', \mathbf{w}),$$

evaluated with:

$$\varepsilon_A = \left\langle H(\mathbf{x}^{ar}_{t\to t+\delta t}) - \mathbf{y}^{o}_{t+\delta t} \right\rangle.$$

Predictability is defined as the time at which a forecast error reaches a background level [Zhang et al., 2019]. Since the errors of the forecasts without assimilation ε_B and ε_O are at their respective saturation levels, we assume them to be the background levels for the forecasts with assimilation. Therefore, in a perfect-model scenario (forecast with initialisation from truth) the intrinsic predictability is the time δt when $\varepsilon_T \approx \varepsilon_B$. The practical predictability is the time δt when $\varepsilon_S \approx \varepsilon_B$. Similarly, in the case of assimilation of real observations the practical predictability is time δt when $\varepsilon_A \approx \varepsilon_O$.

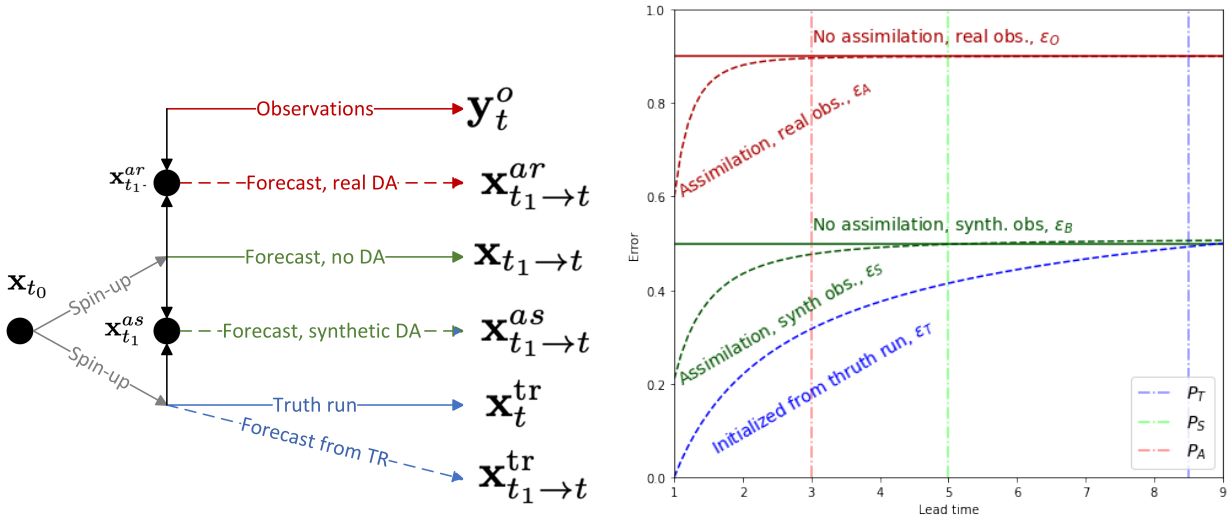

Figure 1: Scheme of experiments (left) and scheme of errors (right). The truth run is shown by solid blue line, observations - by solid red line. The forecasts initiated from the truth - dashed blue line; without assimilation - solid green line; with assimilation of synthetic observations - dashed green line; with assimilation of real satellite observations - dashed red line. Grey lines show spin-up period for the truth and the no-DA forecasts. On the scheme of errors the lines are coloured as follows: red - evaluation against satellite observations, green - evaluation agains synthetic observations, blue - evaluation against the "truth run". Solid lines show climatological level, dashed - average over forecasts. Vertical lines P_T, P_S and P_A indicate potential predictability, practical predictability for synthetic DA and practical predictability for real DA, correspondingly.

Main points

1. According to Fig2 and the text, the data assimilation scheme uses observational data from the period between t0 and t1 to initialise the model at t0; more generally the model is initialised at t(n) with data from between t(n) and t(n+1) when in a realistic system, the data is not yet available. The difference between t(n) and t(n+1) is 24h. Then the same data set is used to evaluate the result of the assimilated model, i.e. the evaluation of the "forecast" on the first day is with the same data as the data set that is used to initialise the forecast. Since in the simple scheme, the initialisation is done neglecting the corresponding models value entirely (weight 1 in Section 3.4, 4.1), this comparison only

shows how well the model persists the initial conditions. Not surprisingly, the model/data "agreement" is quite good on the first day and quickly deteriorates on day 2-5. A proper scheme would use data from t(n-1) to t(n) to initialise at t(n) and compare to data at t(n), t(n+1), etc. As long as this is not change, the first day of the "forecast" cannot be used for any analysis and shouldn't even be called a forecast. The model runs between t(n) and t(n+1).

All the previous and many additional experiments were re-run with assimilation of historical data only. The deformation is now computed from observations of drift at t(n-1) and t(n), then the analysis is computed and the model is initialised at time t(n), then the forecast is compared with deformation computed from t(n) to t(n+1), from t(n+1) to t(n+2), etc. The new experiments confirm that assimilation of data from the day before improves the accuracy of the deformation forecast both in case of synthetic and real data. New results are shown on Fig. 2 and are added to the Results section.

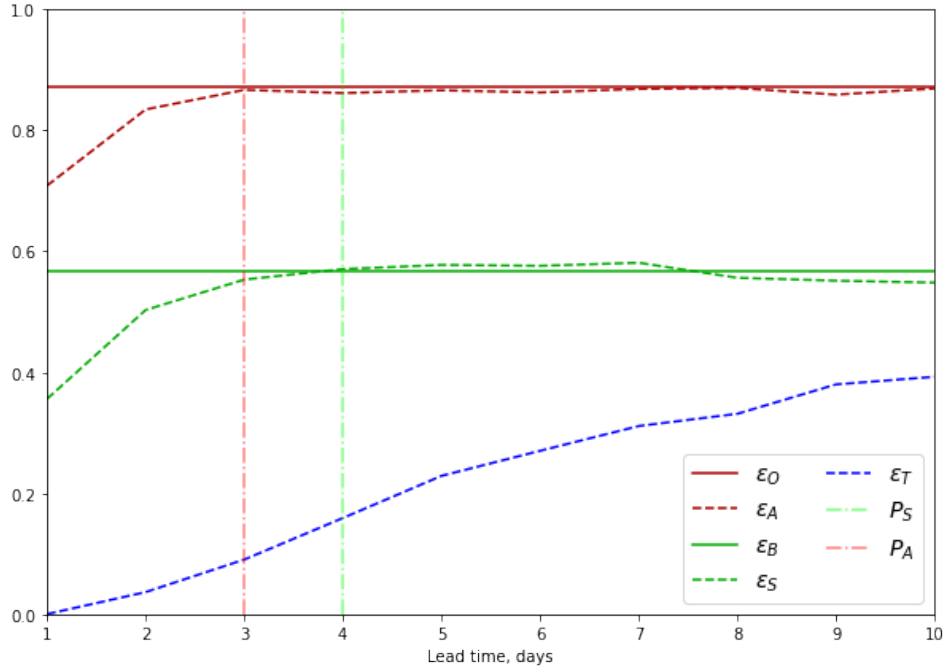

Figure 2: Errors of forecasts. Lines are coloured according to the same scheme as on Fig. 1.

2. Terminology and language. The use of established terminology is rather liberal in the manuscript. As far as I know (but I may be wrong), the terms prediction, predictive skill, predictability, potential predictability have well defined meanings (I haven't heard of "prediction skill"), and the terms should be used when describing the experiments, otherwise it is hard to relate the work to other DA publications. Similarly, the DA scheme is described as "nudging", whereas it is a weighted re-initialisation scheme (according to eq11), but then the weights are always 1 or 0, so there is no weighting in this manuscript. "Nudging" implies a term in an evolution equation d v /dt = some terms + nudgingParamter * (v_o - v), where v_o is the observation. Many smaller problems of similar nature can be found in the text. I marked some of them in the list below.

The terminology has been changed according to the reviewer comments. The term "nudging" is replaced with "direct insertion" [Stanev and Schulz-Stellenfleth, 2014]. In the new experiments we tested different values of the weights [0, 0.5, 1].

3. In data assimilation, one can expect that including additional information will improve the result. Therefore, comparing an assimilated model to a free run makes little sense. Essentially the free run in Fig3 is a 22-day forecast that hasn't seen any new data in 22 days. As noted before, comparing the model to observations that have been used in the assimilation cannot say much about the "success" of assimilation. Anything but any small improvement would just be a failure. Similar "mistakes" have been made before, e.g. doi:10.3189/2015AoG69A740
In general, one would have a ctrl-simulation with an established DA scheme and then add new data or new methods and compare the improvements over the ctrl-simulation. This is done in section 4.2,

but it would be more interesting to see, if the addition of deformation data to an existing DA system (which may already assimilate ice concentration or even thickness) would improve the predictive skill. The authors present their work as a "proof of concept", but the evidence they provide does not help in evaluating, if these additional data help in a realistic system, because the framework is so different.

In the new experiments we evaluate the forecasts on independent data which was not used for assimilation. As explained above, the difference between the forecasts without assimilation and the observations provides the background error. The error of the forecast with assimilation is compared to the background error for evaluating practical predictability. Our previous experiments with assimilation of sea ice concentration [Williams et al., 2021] showed that ice drift is not significantly affected by assimilation. In that sense the ctrl-simulation with assimilation of, e.g., ice concentration is almost equal to the forecasts without assimilation.

4. When introducing new data and constraining new model variables it is good practice (also in sea ice data assimilation) to test schemes and types of data in twin experiments, where a free run produces "observations", a subset of which is used for assimilation leaving the remaining data for validation. This has been done a lot especially in anticipation of new data (e.g. doi:10.1029/2006JC003786). Here, one could have at least held back some of the observations to be used for model validation. This makes it impossible to check if the DA actually improves the state also away from the observations. Instead, one can only make statements about the plausibility of the solutions outside the areas covered by observations (by no mean can the authors claim, that the LKFs are "corrected" outside of the data coverage).

The following twin experiments were run as explained above:

- Truth run,
- Forecast initialised from truth,
- Forecast without assimilation,
- Forecast with assimilation of synthetic data,
- Forecast with assimilation of real data.

As explained on Fig. 1 and 2 these experiments allowed to evaluate the practical predictability on synthetic observations and real observations.

Satellite observations already have large gaps and artificially reducing the coverage even further will decrease the ability of neXtSIM to connect the pieces of LKFs. Luckily we have a few cases when the coverage of test data exceeds the coverage of assimilated data, and we can test our hypothesis that LKFs are improved also outside the area of assimilation as shown on Fig. 4 in the manuscript. In the new set of experiments we use independent data for testing and can confirm that MCC increases also outside the area of assimilation. Nevertheless, due to lack of sufficient cases of confirmed good extrapolation, the statement on correction of LKFs in the entire basin is rephrased as follows:

"... it illustrates that even if data insertion is spatially limited by satellite observations (or even very localized in high deformation zones) it can realistically extrapolate the deformation pattern by connecting the elements of linear kinematic features."

Experiments with spatially limited assimilation of synthetic data also confirm that accuracy of LKFs is increased not only in the area of assimilation, as shown on Fig. 3.

5. A key point of the procedure is how the deformation derived from satellite ice drift data is connected to the model variables concentration and damage. The derivation of this empirical connection is moved to the "supplemental material/Appendix I", which is not part of the manuscript, nor can it be found online.

Appendix I was not included in the first submission due to a technical problem. It will be included in the manuscript.

Minor issues, typos, suggestions, some related to the above points.

page 1
l2: due to the lack $-$ > in the next sentence there are observations to be assimilated. Please rewrite.

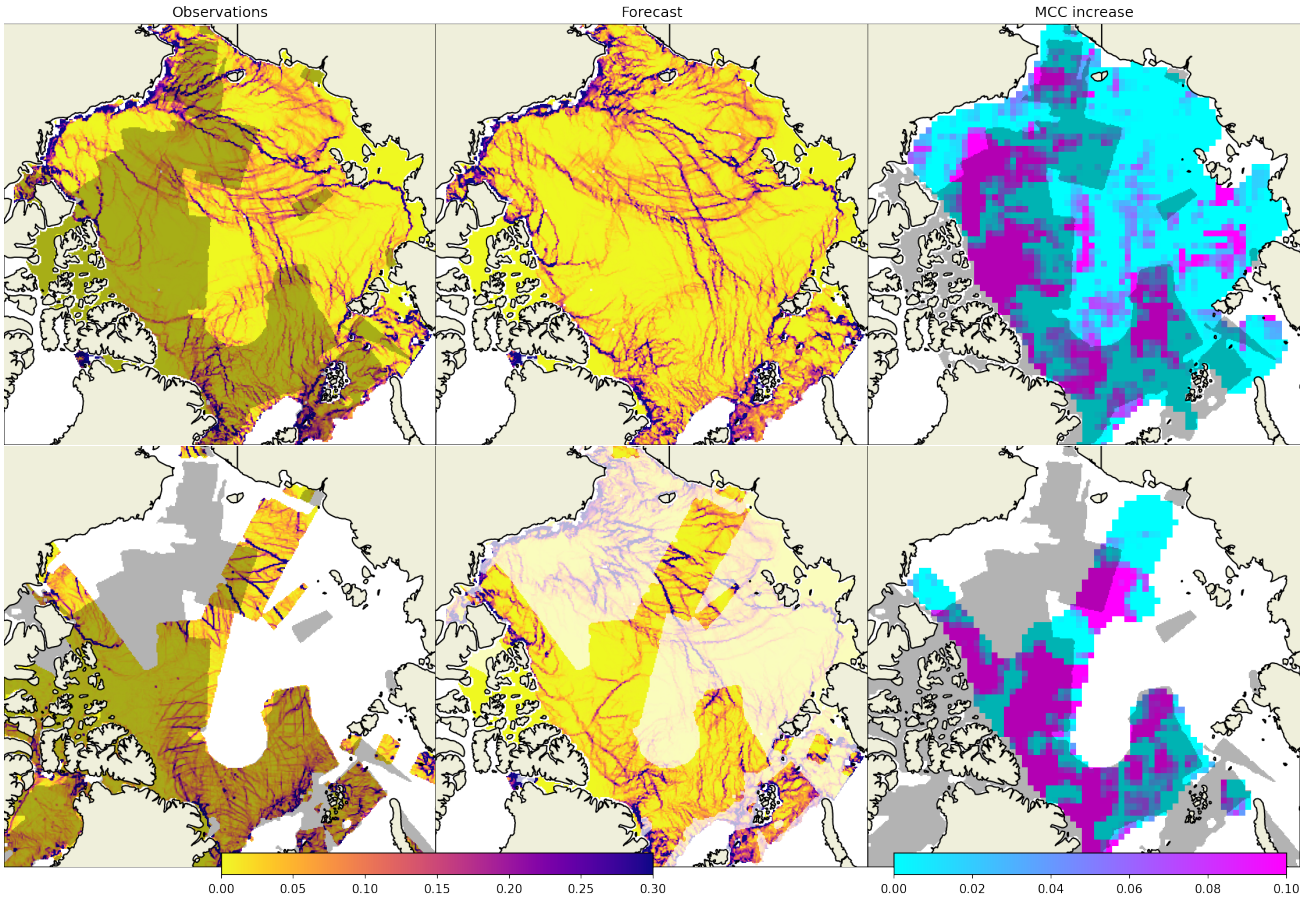

Figure 3: Impact of synthetic (upper plots) and real (lower plots) DA outside the region of assimilation. The semi-transparent grey mask shows area where data from 22 January was assimilated. The "Observation" and "Forecast" maps are for 23 January. The "MCC increase" map shows the increase in MCC compared to the forecast without assimilation.

Rewritten as follows: "Short-term sea ice predictability is challenging despite recent advancements in sea ice modelling and new observations of sea ice deformation, that capture small-scale features (open leads and ridges) at kilometre scale."

l8: deterministic forecasting with one member − > isn't that a homoioteleuton? A one-member system is always deterministic, or an ensemble with just one member, is not really an ensemble. Remove "with one member", or replace "with a single simulation"

The sentence is rewritten as follows: "The proof-of-concept assimilation scheme uses a data insertion approach and forecasting with one member. We obtain statistics of assimilation impact over the long test period with many realisations starting from different initial times."

l10: in 3–5 days horizon − > grammar?

The sentence is rewritten as follows: Assimilation and forecasting experiments are run on synthetic and real observations in January 2021 and show increased accuracy of deformation prediction for the first 2 - 3 days.

l13: reduction in: article missing, or replace by "reducing the", although this sentence is not very clear in general and could be improved

The sentence was removed.

l13: bigger role − > than what?

The sentence was removed.

l20: only − > mainly (sea surface tilt, momentum advection, and you cannot exclude small effects of floe-floe interaction) or "dominated"

The sentence is rewritten as follows: "...the speed and direction of the drift are dominated by the atmospheric and ocean drag forces and by the Coriolis force."

l23: brittle $->$ I don't think that you can say that. It's driven by complex non-Newtonian mechanics/dynamics, but "brittle" is just one aspect of it, and frankly, only a model for the behavior. Other models of sea ice motion exist (I don't mean numerical models). Please rewrite.

Relation to brittle rheology is removed.

page 2, l25: deforming $->$ just to illustrate my previous point, this deformation is NOT brittle, but plastic (no restoring force pushes the ice back into the initial state as for elastic behaviour). The brittle part is just the way failure is parameterised in nextsim (which I believe is a good model for this). I think that this general description of sea ice mechanics/dynamics needs to be "decoupled" from the specific model nextsim, that is being used in this manuscript.

The simplified explanation of brittle rheology is rewritten as follows:

"First, under increasing external forcing the undamaged ice deforms primarily as an elastic material. Internal stresses gradually accumulate in the material until a failure criterion is reached, which corresponds to a limit when sea ice fractures, and then the ice starts deforming along the multiple narrow and elongated cracks, and does so until these later refreeze or when the load (winds and currents) on the ice changes."

We agree that brittle sea ice deformation is more general than a specific BBM rheology or a specific implementation in neXtSIM. In the paragraph near line 25 neXtSIM is not mentioned.

l28: "Under divergent ice motion these cracks become open leads, significantly increasing ocean-air heat and mass exchange and modifying local atmospheric boundary layer and ocean mixed layer. Open leads are also key both for marine fauna survival," $->$ I agree with this, certainly on the local scale of the leads, but this is just a plausibility argument. I have not yet seen that this has been confirmed on large scale heat and mass exchange and budgets. Please give references, if you have them, otherwise marks this as a "plausible assumption".

A relevant reference (Olason et al., 2021) is added.

l37: observe $->$ isn't RGPS a data product derived from Radarsat on a 12.5km grid? I find "observe" in this context inappropriate. Please rewrite.

There are several RGPS products: a gridded deformation product on 12.5 km resolution and a Lagrangian product with a variable mesh size with a nominal resolution of approx. 10 km. The sentence is rewritten as follows:

"The Radarsat Geophysical Processor System (RGPS) dataset was the first attempt to systematically observe sea ice drift and derive sea ice deformation on high spatial resolution (10 km) and with high frequency (3 days) over a long period of time (winters 1996–2016) (Kwok, 1998)."

l41: (10 - 30 km) $->$ can be only a result of the "coarse" resolution, or do we really have "cracks" in the Arctic that are 10-30km wide. Those would be large stretches of either open water or vigorous deformation. I assume that the interpretation is important for the DA.

Rewritten as follows:

The cracks appear on satellite-derived ice deformation products as narrow (10 - 30 km, depending on resolution of satellite data) and long (up to 1000 km) lineaments and are also called linear kinematic features (LKFs) (Kwok, 2001).

l44: "only one model, neXtSIM" $->$ Supposedly this is put here to justify the decision to use neXtSIM. The statement is not incorrect, but it is not clear to me, what the authors would like to achieve with this statement. It does not help this paper in any way, because it hides the results of the Bouchat's paper, that other models have similar properties (at finer grid spacing and higher computational cost). Also doesn't the nextsim in Bouchat's paper use the MEB rheology instead of the BBM-rheology? I would rewrite as something like this (I tried to emphasise that this is a useful model for this study, i.e. does the job very well and is comparatively cheap):

In a recent model intercomparison paper (Bouchat et al., 2021), neXtSIM simulations (neXt Generation Sea Ice Model, Bouillon and Rampal, 2015a; Rampal 45 et al., 2016), ranked among the best for simulating the observed probability distribution, spatial distribution and fractal properties of sea ice

deformation, even though it operates on a low resolution grid of 10km. All other comparable simulations used higher resolution and were hence more expensive.

As pointed out by the second reviewer, some of the models were running at a comparable resolution with quite good results. Therefore, the text is rewritten as follows:

"In a recent model intercomparison paper (Bouchat et al., 2021), neXtSIM simulations (neXt Generation Sea Ice Model, Bouillon and Rampal, 2015a; Rampal et al., 2016), ranked among the best for simulating the observed probability distribution, spatial distribution and fractal properties of sea ice deformation, even though it operates on a low resolution grid of 10km. Analysis of spatial and temporal scaling (Fig. 13 in Bouchat et al., 2021) shows that the spatial structure function of neXtSIM matches the RGPS observations very well, whereas the temporal one is overestimated by 3 - 5 %, probably indicating overestimation of the intermittency by neXtSIM."

l49: skill $->$ the skill?

Corrected

l51: observations $->$ the technical term for this is "potential predictability", which always excludes observations. Why not use that?

Rewritten as follows:

"Mohammadi-Aragh et al. (2018) evaluated the potential predictability of LKFs using an ensemble of sea ice models all using a viscous-plastic rheology, but practical predictability remains unknown."

l55: "so the assimilation scheme needs to perform a cross-variable update from deformation to sea ice model variables." This is a common "problem" in DA and one would use a proper "observation operator" that maps the model variables to the observations. The dual operation then maps the model-data misfit back to increments of model variables. If you want to talk about "data assimilation", I suggest to use the proper language/terminology. Here you will (according to the abstract) do a nudging experiment (but it turns out to be re-initialisation in reality), which is strictly speaking not really data assimilation (although total valid as a method).

The term "nudging" is replaced with "direct insertion", which, according to e.g. [Stanev and Schulz-Stellenfleth, 2014], is one of the data assimilation methods. The sentence is rewritten as follows:

"First, the "direct insertion" method operates in the model state space. However, the observed deformation is not a model prognostic variable, so an operator is used to convert deformation to the model variables. This operator is an inverse of the observation operators used in data assimilation, since it maps from the observation space back to the state space."

page 3: l79: "d is the ice damage." maybe it makes sense to clearly state the mean of "d", d=0 is entirely intact and 1 is entirely damaged (which I assume here from the equation) or vice versa. Previous publications use contradicting definitions.

Explanation is added: "(with $d = 0$ being completely undamaged ice)."

page 4: eq5, what is "P"?

Added after Eq 5.:

"where P is a constant scaling parameter for the ridging threshold to parameterise P_{max}, following the results of Hopkins (1998), and h is thickness."

l101: "12 hours frequency" $->$ 12h is not a frequency, but a period. The frequency is: one record in 12h.

Rewritten as follows:

"Ice drift is computed from pairs of images separated by approximately 24 hours and the product is delivered every 12 hours."

page 5, l110: "reach an equilibrium" in 30 days is hard to believe, usually one would expect a seasonal cycle at least, unless the sea ice models of TOPAZ4 and nextsim are identical (which they are not, I assume). But does the equilibrium matter?

The entire description of experiments is rewritten (see above).

Figure 1 is not really necessary.

The entire description of experiments is rewritten (see above) and a scheme with explanations of several experiments becomes necessary.

l113: this looks like the scheme uses data from the future to correct the model? Does that make any sense? I would expect to update the model variable at t(n) with data collected over t(n-1) (or earlier) to t(n). See also main points above.

The experiments were re-run to validate the results on independent data. Please see the replies to the major points.

l119: in the previous work of (e.g. (Bouillon and Rampal, 2015b)). $->$ fix parentheses

Corrected.

Fig2, caption: Eps,d and A $->$ use proper symbols as in figure.

Figure 2 is removed.

page 6: l125 The "Appendix" should be in the same file as this text, right? Supplementary material is separate. What do we have here? On the TC-web page I cannot find any supplementary material, so that "Appendix I" is missing for now.

Unfortunately, the Appendix was not added due to technical reasons. It is added in the revised manuscript.

eq9: that would be $1 - 10^{k_2} * \varepsilon_{tot}^{k_3} - k_1$, right? Now it would be interesting to know at least k_2 and k_3, because that would show how strongly the total deformation impacts damage, compared to eq10, where the impact is linear but later in the full equations exponential, as argued in the discussion section 5.2

These values are provided in the Appendix.

eq10: wouldn't i make sense to treat divergence and shear separately, ie. have two different coefficients: $f_A = 1 - a_1\varepsilon_{div} - a_2\varepsilon_{shear}$, or even differential between divergence and convergence. It is clear the divergence will create open water directly, but convergence will do this to a much smaller degree (e.g. lateral divergence in convergence), and also shear should have a different coefficient.

We assume that all deformation events (convergence, divergence and shear) indicate presence of weaker ice that may continue to be deformed. Ice weakness is simulated in neXtSIM by decreased concentration or increased damage. Observation of any of deformation components (including convergence) is interpreted in the assimilation procedure as an increase in ice weakness and, therefore, decrease in concentration or increase in damage. We cannot find reasoning why weakening of ice due to convergence is different from weakening due to divergence or shear and, therefore, suggest that total deformation is a good proxy for detection of weak ice and a single dependence of A (and d) on total deformation can be used. Corresponding explanations are added to the text.

page 7 l148: simple least-squares nudging approach? Where are the least squares? Deriving eq11 from a least-square formation is possible but a little vain. I am not sure if I would call eq11 "nudging", as nudging usually implies a time varying equation such as $dv/dt = rhs + nudginpgarameter * (v_o - v)$, which is not what eq11 implies. If in a DA cycle $v_m(n)$ is computed at time t(n), then updated according to eq11, and then $v_a(n)$ is used to initialise the next DA cycle, then this is not nudging, but re-initialisation with a very simplified updating scheme. I am not criticising that, but I think that the description needs to be accurate. See also main comments.

Rewritten as :

We update the damage and concentration variables in the model according to the observed deformation using a "direct insertion" method (Stanev and Schulz-Stellenfleth, 2014) as a proof of concept for DA.

l159: "the very small spatial correlation approximation is reasonable." I disagree. This assumption is valid normal to the fracture, but along the fracture, considering the nearly instantaneous fracture propagation (in nextsim and in observations), this not a "reasonable" assumption. Please rephrase.

Rephrased as:

"Considering that sea ice deformation is accommodated along nearly 1D geometrical features (i.e. fractures), some correlation can be seen only along the fraction, and approximation of a low spatial correlation in all other directions is reasonable."

l164: value of eps_min? Mention here, that this is part of the sensitivity analysis?

Rewritten as:

"where ε_{min} is a threshold for total deformation found in sensitivity experiments."

l171:"Since it was difficult to distinguish between the individual impacts of w_v and W in Eq. 12,", unclear, why.

The number of experiments was increased, and values of 0, 0.5 and 1 were tested. The sentence was removed.

l172: "0 and 1 were tested for w_d and w_A", but this means that there is no weighted average at all and all that is done is re-initialisation. I think it would help the reader to clarify the scheme: Either, there is pure re-initialisation with a somehow derived value, or no re-initialisation. The entire description of least-square nudging is entirely misleading (and does not describe, what is actually done). See main comments

The number of experiments was increased, and values of 0, 0.5 and 1 were tested. The sentence was removed.

page 8, l184: "difference in 90th percentile", what is that? Please be more specific.

The 90th percentile is not used anymore for estimating predictability as it shows very similar results to A_{MCC}. In the sensitivity tests it is replaced with the Kolmogorov-Smirnov test applied to PDFs of forecasted and observed deformation.

l191, related to l184. From the explanation is it no clear what is computed here, and in which sense this is different from "MCC". Is MCC a standard statistical method, or something that is only described in Korosov and Rampal, 2017? If it is a standard method, please cite a standard reference/textbook.

Explanation of the MCC computation is added to the Appendix. A reference to a relevant textbook (Brunelli et al., 2009) is added.

Further, there were a few metrics suggested in the cited papers by Bouchat et al 2022, and companion paper Hutter et al 2022, also Mohammadi-Aragh et al 2018. In what sense are the metrics used here related, or do they quantify entirely different properties?

MCC is slightly different to the LKF metrics used in Hutter et al 2022 and Mohammadi-Aragh et al 2018. The following explanations are added:

"Unlike the LKF evaluation metrics suggested in (Hutter et al, 2020) that compare only statistical properties of LKFs (number, density, length, orientation, etc), the MCC metric estimates co-alignment of individual LKFs on model simulations and satellite observations. It is also thought to be more sensitive to LKFs with low deformation magnitude, as no threshold is applied for their detection."

l195: "22nd January 2021" depending on the definition of "free run", I would expect that after 22 days of integration the "free run" has already quite deviated from the observations. What is the point of this comparison? Showing that the model can be "kept on track" even with simple methods, compared to not doing anything? Normally in DA, one defines a baseline/ctrl with some existing system (not the free run!) and compares how the details of the algorithm affect the solution, as has been done in section 4.2. It would also be interesting how important the observations on the current day are for forecasts, i.e. comparing a run with DA until t(n-1) to a run with DA until t(n) (where, in fact, the observations are not from one day into the future as is the case here). See also main points

As explained above the "free run" has never seen observations of deformation and can be considered as a control simulation. The free run is not expected to match with observations (even if spatial patterns of deformation are remarkably similar). The goal of the comparison is to show that the suggested assimilation puts the model "on track", i.e. the forecasts with DA start to match with observations much better than the free run (without DA or with assimilation of something other than deformation). These aspects are also covered in replies to the major points.

Highlight, page 9, l209: "due to its rheology" $->$ since only nextsim is used with one rheology, this (part of the) statement is not supported by the experiment and should be removed.

Relation to rheology is removed.

Also, since there is no observational data to check the results in the "unobserved" regions, one cannot claim that "neXtSIM is able to extrapolate and create realistic connections". The model simulation creates connections that look realistic, in the sense that they are not garbage, but that's about it. For a statement like the one in ll209/210, one needs experiments where part of the data is withheld from the DA, to be used later for model validation.

As explained above, in the new set of experiments we use independent data for testing and can confirm that MCC increases also outside the area of assimilation. Nevertheless, due to the lack of sufficient cases of confirmed good extrapolation, the statement on correction of LKFs in the entire basin is rephrased as follows:

"... it illustrates that even if data insertion is spatially limited by satellite observations (or even very localized in high deformation zones) it can realistically extrapolate the deformation pattern by connecting the elements of linear kinematic features."

Further, in DA we expect that the results improve with additional data. Any other result would be failure of the DA, so all that figure 3C shows, is that the DA algorithm does, what is has been designed to do. This comparison is even further biased, because now (according to the description in Section 3, Fig2) the model has been corrected with data from the future (t(n)+24h), and then is compared to the same data from the future. I wouldn't call that prediction, but analysis.

As explained above, in the revised manuscript observations from t(n-1)-t(n) are used for analysis on t(n) and the forecast on t(n)-t(n+1) is evaluated on independent data from the same period. The goal of Fig. 3 is exactly to illustrate that the DA algorithm does what it is supposed to do.

page 11, Figure 4, please add colorbars to make it easier to view the images

The colorbars are added.

page 12, l218: persistent or persistency

The presidency forecast is not used anymore in evaluations according to reviewers requests.

l226: fix D_P90

Corrected.

l227: assimilation, better : re-initialisation.

Rewritten as "assimilation using direct insertion"

l233: sufficient − > a sufficient

Corrected.

l235: consequent − > subsequent

Corrected.

l238: nudging − > re-initialisation

Rewritten as "That demonstrates the impact of insertion of excessive concentration decrease, ..."

page 13, l242: "The experiments with w_d cannot detect" anything, rewrite as "In the experiment with w_d, one cannot detect ..."

Rewritten as suggested.

Figure 6 tells me, that the leading order effect is achieved by "a_1", (except of a_1=-2, where a similar effect is achieved by modifying eps_min), so the linear relationship between total deformation and ice concentration. All other parameter appear to have small effects only. Maybe this should be stated somewhere explicitly.

This observation is added to Section 4.3

Fig6 is difficult to read, maybe make the bars broader?

Figure 6 is re-plotted with results from more experiments.

In Tab1 a_1 parameters are all positive, here, they have a negative axis, please correct, also there's seem to be experiments with a_1¡0 (i.e. to the right of 0), which are not listed in Tab1

The sign of a1 on Figure 6 was incorrect and is corrected in the revised manuscript.

l246: "first successful attempt", What is meant by "successful" here? This sounds like a conclusion that needs be backed with evidence. Also since the observations assimilated appear to come "from the future", the results for the first day (which is most "successful") cannot be used.

The observations from future are not used anymore. Fig.5 evidences the success. The sentences is rewritten as follows:

"We present the first successful attempt to use the observed sea ice deformation for increasing accuracy of deformation prediction in the first 2 - 3 days."

l250: "it" − > what is "it"? The relationship between deformation fields and model state variables is not shown by the DA, but by a prior correlation analysis, which I cannot evaluate, because it is moved to an appendix/supplemental material that is not accessible at the moment. Also the damage assimilation had little effect, so that questions both the empirical relation in eq9 and/or the "success" of the DA.

Appendix is added and the sentence is rewritten as follows:

"Our study demonstrates in practice that information contained in the observed deformation fields can be used for initialisation of model state variables."

l252: "proves" − > this is clearly too strong.

Rephrased as follows:

"Third, it illustrates that even if data insertion is spatially limited by satellite observations (or even very localized in high deformation zones) it can realistically extrapolate the deformation pattern by connecting the elements of linear kinematic features."

l253 "corrects" − > to correct means to make it right, but there's no proof for that in the manuscript. All that the experiments show is that the model takes the initialisation information and propagates it sensibly (according to the model dynamics) into areas that have not been re-initialised. This does not mean that we now have "correct" forecasts, just that there's some "dynamical extrapolation" that needs to be evaluated with independent data (and this important step it is missing). See also main points.

See previous reply.

page 14, l266: this paragraph sounds like a project proposal with some selling arguments. Not sure if a scientific publication is the right place to advertise one's work in such a way. In my view, a scientific publication in TC should report scientific advances, but not the suitability of a system for tasks that have not yet been performed. Please rewrite or remove.

The paragraph on practical usefulness is removed.

page 16, l302: skill for 2–5 days − > see earlier comments, I think that the first day cannot be counted because of the data from the future.

As seen from Fig. 3, without using data from future the accuracy of forecasts is improved for 2 - 3 days.

l307 to the end of the section: I think that this list of factors impacting the predictive skill of LKFs would be much better placed (slightly modified) in the introduction, to lay out the scope of the manuscript and which of these aspects will be addressed in the manuscript.

The list of factors impacting predictability is added to the introduction, but a more detailed description remains in the Discussion section as it leaves many open questions for future research.

l325: Bouillon et al., 2009 − > wrong reference. The correct Bouillon paper is from 2013, where this is called "revised EVP", although I believe that the proper reference would be Lemieux et al 2012, who were the first to modify EVP which then was described as modified EVP in Kimmritz et al (2015). It is not clear to me, how using a VP rheology (mEVP is a method to solve the VP rheology equations),

that has been marked as too slow, etc. in this paper and many other papers of this group, is going to help here at all.

The reference is corrected. The following clarifications are added:

"We expect that the model equipped with the mEVP rheology will not be capable of spatial extrapolation of the assimilated ice weakness (lowered A or enhanced d), and that further tuning of the BBM rheology can improve the practical predictability of LKFs."

page 17, l349: "neXtSIM is capable of extrapolating the spatially discontinuous satellite observations of deformation by connecting the elements of linear kinematic features in a realistic manner." − > this is a statement, that I think is totally justified from the evidence provided (Fig3). Please rewrite previous statements about "correcting" LKFs etc accordingly.

The statement is rewritten (see above).

l351: local − > locally?

Corrected.

page 18 l359: Data availability: TOPAZ data and other forcing data are not mentioned, no code availability.

The Data availability section is updated correspondingly.