

Author's response to Referee 1
for 'Four-dimensional variational data
assimilation with a sea-ice thickness emulator'

Charlotte Durand, Tobias Sebastian Finn, Alban Farchi,
Marc Bocquet, Julien Brajard and Laurent Bertino

July 2025

RC: Reviewer Comment; **AR:** Author Response

RC: The manuscript "Four-dimensional variational data assimilation with a sea-ice thickness emulator" by Durand et al presents an evaluation of a 4D-Var data assimilation framework using a data-driven sea ice thickness emulator of the neXtSIM sea ice model. The authors show how, through the emulator's back-propagation capabilities, sea ice thickness observations (both idealized and real) can be assimilated into the emulator using 4D-Var, effectively reducing the emulator bias. I very much enjoyed reading this manuscript and think it's a nice contribution to the literature. In fact, most of the comments I had noted down by the time of the discussion were then answered in the discussion, so thanks! My comments were overall minor, and I think the manuscript is almost ready for publication with a few small edits (see below).

AR: We deeply appreciate the reviewer's thorough and insightful review of our work. In the following, we respond to the comments and raised issues and point to the changes in our manuscript.

RC: General question regarding methodology

RC: Could you just clarify something about the methodology for me. Are the EOFs used for the background covariance static over the course of the DA simulation? My concern early on was the ability of the EOF approach to capture flow-dependent processes, given the strong seasonal cycle of sea ice (you do mention this

in the discussion). Is there some expectation that the minimization figures out which EOFs are most important and dynamically weights them (in time) according to w ? I would be very interested to see how the approach compares to an Ensemble Kalman Filter (as you also say in the discussion).

AR: Thank you for your remark. Indeed we investigated the influence of the different EOFs weights and their seasonality although we did not present the results in this paper. They are presented in Durand (2024) in Chapter 7. Regarding the 4D-Var-EOF, we can assess the significance of the different weights associated with the EOFs and evaluate the time dependency of the predominant ones. The results are presented in Fig. 1 of the present document. Firstly, the weights associated with the largest amplitudes correspond to the first EOF coefficients. Secondly, tracking the time evolution of the first coefficient reveals a clear temporal dependency, in line with the annual evolution of the SIT. The second coefficient also exhibits a seasonal behavior, with an increase in amplitude around May. The 4D-Var-EOF approach captures the seasonal variability of the signal, which is expected to be the dominant source of temporal variability. However, the full flow-dependent covariance structure, dependent not only on the season but also on the specific realization of the forecast, cannot be represented by 4D-Var-EOF. Capturing this would require comparison with an ensemble-based method such as the EnKF, which would be interesting but is beyond the scope of this article. Note that a standard EnKF system does not rely on an EOF decomposition. However, if the ensemble were projected onto the same EOF basis, it would be possible to compute time-evolving weights w from the EnKF that vary with both location and time, making a direct comparison impractical.

RC: Comments

RC: L112: I suggest adding a citation to show an example of where observations are typically log-normal. E.g Landy et al 2020.

AR: Thank you very much for your suggestion, we added this reference: "as more commonly encountered in sea-ice observations from satellites (Landy et al., 2020)"

RC: L117 and elsewhere: change "In average," to "On average,"

AR: Thank you for seeing this, we corrected it and checked thoroughly the rest of the paper.

RC: L129 - L132: Somewhere in this section it might be worth high-

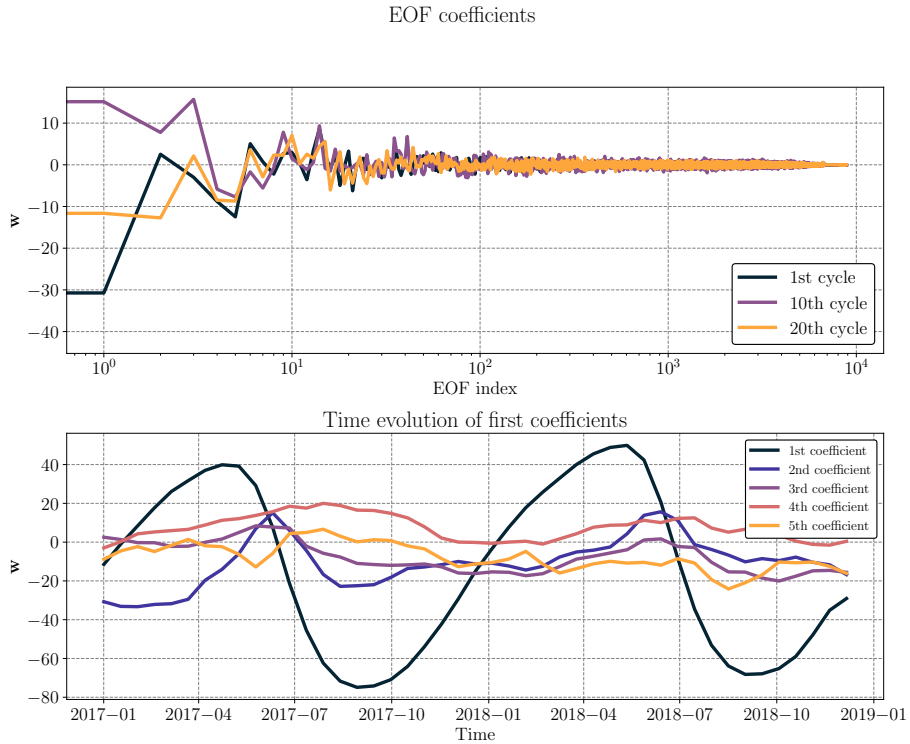


Figure 1: Upper panel: Values of the EOF weights for different assimilation cycles: the 1st cycle is represented by the blue line, the 10th cycle by the purple line, and the 20th cycle by orange line. Note that the x-axis is on a logarithmic scale to highlight the first weights. Bottom panel: Time evolution of the weights of \mathbf{x}_a associated with the first five EOFs for each cycle. The first coefficient is shown in dark blue, the second in indigo, the third in purple, the fourth in pink, and the fifth in orange.

lighting a recent paper (Nab et al. 2025) which quantified the effect on DA-derived analysis fields due to varying observational uncertainty on sea ice thickness measurements—Turns out to be quite sensitive.

AR: Thank you for this paper which we were not aware of. Indeed, this is worth mentioning. We changed the text in L132 to : "Note that Nab et al. (2025) showed that modifying SIT observation uncertainties introduces significant sensitivities during SIT assimilation."

RC: Figures 3 and 7: Missing text in all labels

AR: Thank you for noticing this, we were careful to check if in the PDF the text is lisible.

RC: L258: Doesn't the RMSE in Fig 5 peak in July? I guess the bias error peaks just before May and then rises again in December? Maybe changing L258 to "from Fig. 5 top" to make it clear which panel in Fig 5 we are looking at

AR: Yes thank you for noticing that, the RMSE peak is indeed rather in July, we changed the sentence to "The analysis from Fig.5(top) reveals a strong seasonality in results, with the RMSE peaking in July."

RC: L306-310 : Can you borrow some info from data-driven NWP models which retain sharpness by augmenting loss function

AR: Yes, indeed! For the emulator, we are exploring alternative loss functions to better preserve sharpness. However, these often result in increased RMSEs, so this remains a work in progress and is beyond the scope of the current paper.

RC: L350 : Might be worth highlighting here that there are ongoing developments in this space. For example Chen et al and Gregory et al both show ML-based approaches for deriving complete daily sea ice and ocean fields from satellite altimetry at 5 km grid resolution. Both of these approaches model the spatio-temporal covariance of daily fields, rather than simply averaging through time. Although these studies show sea ice freeboard, it is conceivable that daily sea ice thickness observations are on the horizon.

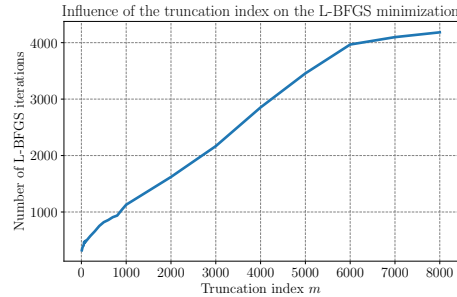


Figure 2: Evolution of the number of L-BFGS iterations as a function of the truncation index.

AR: Thank you for suggesting those references, we added the sentence: "Ongoing works are proposing ML-based approaches to derive complete daily sea ice freeboard fields from satellite altimetry at fine spatial resolution (5 km), by modeling the spatio-temporal covariance of daily fields rather than relying solely on temporal averaging (Gregory et al., 2024; Chen et al., 2024)."

RC: L400: I thought neXtSIM-F was initialized through nudging and not EnKF (L274/275)?

AR: Thank you for seeing this error, neXtSIM-F is indeed initialized with nudging operationally, we corrected the sentence: "With limited resources, such as emulating only sea-ice thickness and assimilating CS2SMOS observations, the developed 4D-Var system performs comparably to the operational neXtSIM-F system"

RC: Appendix B: Can you quantify the time change in the 4D-var minimization when increasing the truncation index m ? For example, on L324 you say it's 155 seconds for $m=7000$. What is the time if m is halved to 3500? I guess I'm wondering what is the cost-accuracy tradeoff.

AR: We show in Fig. 2 of the present document the number of L-BFGS iterations as a function of the truncation index. The execution time to run the 4D-Var is proportional to the number of iterations. As you can see, the increase of the number of iterations is quite linear with the truncation index. When using this method in operations, the choice of the truncation index would be an important factor for the computational efficiency. Yet, to obtain the best RMSE values, based on our experiments, choosing a value above $m = 5000$ seems preferable.

References

- Chen, W., Mahmood, A., Tsamados, M., and Takao, S. (2024). Deep random features for scalable interpolation of spatiotemporal data.
- Durand, C. (2024). Deep learning, data assimilation and sea-ice dynamics.
- Gregory, W., MacEachern, R., Takao, S., Lawrence, I. R., Nab, C., Deisenroth, M. P., and Tsamados, M. (2024). Scalable interpolation of satellite altimetry data with probabilistic machine learning. *Nature Communications*, 15(1).
- Landy, J. C., Petty, A. A., Tsamados, M., and Stroeve, J. C. (2020). Sea ice roughness overlooked as a key source of uncertainty in cryosat-2 ice freeboard retrievals. *Journal of Geophysical Research: Oceans*, 125(5).
- Nab, C., Mignac, D., Landy, J., Martin, M., Stroeve, J., and Tsamados, M. (2025). Sensitivity to sea ice thickness parameters in a coupled ice-ocean data assimilation system. *Journal of Advances in Modeling Earth Systems*, 17(3).

Author’s Response to Referee 2
for ‘Four-dimensional variational data
assimilation with a sea-ice thickness emulator’

Charlotte Durand, Tobias Sebastian Finn, Alban Farchi,
Marc Bocquet, Julien Brajard and Laurent Bertino

July 2025

RC: Reviewer Comment; **AR:** Author Response

RC: This manuscript assesses the performance of 4D-Var data assimilation of sea-ice thickness (SIT) in the context of an SIT emulator featured in a recent publication of *The Cryosphere* (DOI: 10.5194/tc-18-1791-2024), taking advantage of the easy availability of the emulator’s adjoint. It is a scientifically interesting development that, I believe, is worth having its place in the scientific literature. Therefore, having considered the manuscript’s contents in relation to the remit of *The Cryosphere*, I recommend its publication in the journal in principle. That being said, I do have major reservations about the manuscript in its current form, and I think it needs to be substantially revised before publication. The reasons are given in the “Major Comments” section below and elucidated in more detail in the “Minor Comments” section. I understand the list of minor comments is very long, but I hope they could provide concrete suggestions on how the manuscript could be improved, which will also help address the major comments.

AR: We deeply appreciate the reviewer’s thorough and insightful review of our work. In the following, we respond to the comments and raised issues and point to the changes in our manuscript.

1 Major comments

RC: Overall, I find the Introduction (Section 1) a bit fragmented and lacking a general theme or direction that the authors would like to guide the reader towards.

AR: Thank you for your remark. We revised the introduction to make it less fragmented and reorganize the paragraphs to improve the overall flow and readability.

RC: For better clarity, I recommend a reorganisation of the manuscript, to avoid the reader having to go back and forth between the idealised-observation experiments and the CS2SMOS experiments. My recommendation is as follows. After the introductory section, the model, the emulator, the two flavours of 4D-Var and the metrics used for evaluation would be presented first (possibly in separate sections). Next, there would be two main sections, one about the idealised-observations experiments, and the other about the CS2SMOS experiments. In each of these sections, you could first describe the setup of the observations before moving on to discuss the experimental results. Following these sections, some discussion relevant to both sets of experiments could be presented in the final section, where concluding remarks could also be given.

AR: Thank you for your proposition. We reorganized the results section, splitting it in two, and adding to each the definition of the observations. We also reorganized a bit the discussion part.

RC: I think there is too much content in the appendices. Some of the contents are important and should sit in the main text (see e.g. Minor Comments 47 and 104). Other parts of the appendices (such as Appendix E) could be scientifically interesting, but I don't see a good reason that the material should be included in the first place (cf. Minor Comment 72). There should be good coherence between the main text and appendices.

AR: Thank you for raising this point. We have justify our choices in the following Minor Comments (72, 104). We decided to move the section on observation time frequency in the main part and keep the truncation index in the appendix and emphasized in the main text where they can be found. Additionally, we have chosen to retain Appendix E in the manuscript. As you mentioned, it is scientifically interesting and raises an important point, particularly in the discussion, even though the results have not necessarily been evaluated with the same depth as in the main part of the manuscript.

RC: A recurring theme in the minor comments below is that results or claims may not be adequately explained or backed by evidence. Generally speaking, statements need to be more qualified, and caveats need to be mentioned. This is to avoid the impression that the results are over-generalised.

AR: We thank the reviewer for their insightful questions and have provided our responses to the best of our ability.

RC: Another recurring theme in the minor comments below concerns the language and presentation. Despite the manuscript having been grammatically edited with ChatGPT (cf. line 555), there remain a number of grammatical errors and many instances where the usage of the English language could be improved. Some abbreviations and symbols have not been defined at their first use, and some parts of the text may sound a bit cryptic to a non-specialist audience. Please revisit the text and make it more accessible to a general audience.

AR: Thank you for your valuable feedback. We have carefully addressed your comments and incorporated additional definitions and abbreviations. Please see the following minor comments, where several definitions and variables have been properly introduced.

2 Minor comments

RC: 1. Line 21: Please explain what neXtSIM is.

AR: Thank you for your remark, we moved the neXtSIM description (Lines 33 - 34) alongside its references to Line 21: 'This emulator aims at reproducing the evolution of the sea-ice thickness as modeled by the state-of-the-art sea-ice model neXtSIM (Rampal et al., 2016; Ólason et al., 2022)'

RC: 2. Line 24: CryoSat should be CryoSat-2; SMOS has not been defined.

AR: Thank you, we changed Cryosat to Cryosat-2. We defined SMOS abbreviation and changed the sentence to: 'with the assimilation of SIT retrievals from the merged product of Cryosat-2 altimeter and the Soil Moisture and Ocean Salinity (SMOS) radiometer, called CS2SMOS'

RC: 3. Line 25: What does the “-F” refer to?

AR: The F refer to forecast. As the sentence already mention that it is an operational forecasting product, we simply added the following mention: '... but for operational forecasting, the neXtSIM-F system resorted instead to a simple nudging technique (Williams et al., 2021), with -F standing for "Forecast"'

RC: 4. Lines 26 – 32: You open the paragraph using the term “data assimilation techniques” and later you mention about nudging. However, I wouldn’t call nudging a DA technique. Perhaps “initialisation techniques” would be a better term.

AR: While discussable, nudging is often included in data assimilation techniques (Lakshmivarahan and Lewis, 2013). Typically, Lindsay and Zhang (2006) define its nudging as a data assimilation technique. However, we

slightly changed the wording of the sentences to express this: 'Other techniques use covariances that are static in time such as optimal interpolation (Wang et al., 2013; Ji et al., 2015), or three-dimensional variational assimilation (Hebert et al., 2015; Toyoda et al., 2015; Lemieux et al., 2015) and are employed in diverse systems (Caya et al., 2010; Donlon et al., 2012; Zuo et al., 2019). Initialization techniques using nudging (Lindsay and Zhang, 2006; Tietsche et al., 2013) have also been employed.'

RC: 5. Line 27, regarding the word “predominantly”: I am not sure if `enkf` is the most popular way of initialising sea ice in forecasting systems. Variational methods are also common.

AR: Thank you, we changed "predominantly" to "often".

RC: 6. Line 41 – 42: I think the part of the sentence “as well as a background term... state of the system” is a distraction. A background is required for all DA methods, so this is not something special about 4D-Var, yet the way the sentence is worded seems to suggest that this is a characteristic feature of 4D-Var.

AR: Thank you, the structure of the sentence can be improved, we removed the second part of the sentence.

RC: 7. Line 43: The word “at” is unnecessary.

AR: Thank you, we changed the wording to 'The idea behind (strong constraints) four-dimensional variational methods is to estimate a model trajectory that fits best the observations throughout a time period'.

RC: 8. Lines 44 – 45: The use of “to” at the beginning of the sentence seems to suggest that the propagation of gradient information backwards is the aim. However, I think the aim is what you have said in the next sentence instead: to allow information from later parts of the DAW to be incorporated in the analysis at the beginning of the DAW. I think the sentence could be reformulated in a clearer and more concise way.

AR: Thank you, we changed the wording of the sentence: "The propagation of the gradient information from the observational time backwards in time within the DAW relies implicitly on the model's adjoint during the cost function minimization."

RC: 9. Line 45: Why is the minimisation's dependence on the adjoint implicit?

AR: We used the word 'implicitly' because the adjoint is not written in the cost function term explicitly: it is used when the gradient of the cost function is computed by the optimizer. We nonetheless removed the term

in the sentence to prevent confusion: "The propagation of the gradient information from the observational time backwards in time within the DAW relies on the model's adjoint during the cost function minimization."

RC: 10. Line 47: It would be nice to mention the adjoint of NAOSIM (see DOI: 10.1029/2008GL036323).

AR: Thank you, we added this reference. Note that the sea ice model in this coupled model is really simple "The ice concentration is set to 100% where the sea-surface temperature falls below the freezing temperature and an ice thickness of 2m is assumed."

RC: 11. Lines 48 – 49: You say that adjoint models yield limited realism for sea-ice simulations. Yet, as things stand, the remaining sentences in this paragraph don't elaborate on this. I would like to see this explained.

AR: Thank you, we added more details to reinforce this sentence. We meant that a free drift model is not the most realistic model on full Arctic scale. In order to avoid numerical instabilities, strong simplifications are made on the adjoint or the associated sea ice model, typically for the sea ice model of NAOSIM mentioned in the previous point. In Toyoda et al. (2019), they say *Behavior of an adjoint of full sea ice dynamics including rheology has not been documented before mainly because previous adjoint models greatly simplified sea ice dynamics (free drift) to avoid numerical instabilities*. In their effort to have an adjoint for a EVP model, they still simplify some of the adjoint equations (see page 2148 - 2149 of their paper). We changed the sentence to "While adjoints for simplified free-drift models are achievable (Koldunov et al., 2017), they yield limited realism for full Arctic sea-ice simulations. Usui et al. (2016); Toyoda et al. (2015, 2019) developed an adjoint of their sea-ice model, which relies on an slightly simplified elasto-visco-plastic (EVP) rheology scheme (Hunke and Dukowicz, 1997) within a coupled ocean-sea ice model framework to avoid numerical instabilities."

RC: 12. Lines 52 – 53: My understanding is that the adjoint was already available before its incorporation into MIT-GCM.

AR: Thank you for this remark, we changed the sentence to: "Koldunov et al. (2017) studied sea-ice concentration assimilation with an adjoint (Hibler, 1979) in the MIT-GCM model. The rheology in this adjoint is either discarded or simplified." In his case, the rheology is turned off during the adjoint computation to avoid numerical instabilities.

RC: 13. Section 2.1, section heading: This section includes description about adaptations of the neXtSIM grid and model variables to the neural network model you are going to use. Perhaps a section heading reflecting this would be more accurate.

AR: Thank you, we changed the heading to "Physical model simulation and atmospheric forcings preprocessing" (cf Minor comment 17.)

RC: 14. Lines 81 – 82: The abbreviation NEMO OPA is undefined.

AR: Thank you, we added the definition of NEMO, but OPA is just its ocean component. The sentence will become "Additionally, the sea-ice model is coupled with the Nucleus for European Modelling of the Ocean (NEMO) framework's ocean model, OPA"

RC: 15. Line 91: Could you explain why a normalized variable has to be used as opposed to the original variable?

AR: Normalizing values are standard practices in deep learning based methods to stabilize and speed up the training. We also mention that the ERA5 forcings are normalized (Lines 100 - " Interpolated onto the Eulerian curvilinear grid with nearest neighbors, forcings at time t , $t+6h$ and $t+12h$ are normalized and added as predictors to the input of the neural network"). We added a sentence at the end of the paragraph: "The normalization of all input fields in the neural network is a common practice to stabilize and speed up the training (Ioffe and Szegedy, 2015)."

RC: 16. Line 93: When you talk about global statistics (both mean and standard deviation), is the computation of these statistics performed after the coarse-graining, i.e. at the 128×128 resolution?

AR: Thank you for this remark. Yes, we clarified it: "computed over all the coarse-grained grid-cells in the training dataset"

RC: 17. Section 2.1, final paragraph: It may be more appropriate to move this discussion to Section 3 when the surrogate model is described, especially when the forcing vector is mentioned in that section.

AR: We rather modified the heading of the section accordingly "Physical model output and atmospheric forcings preprocessing". To keep all of the data introduction in a same section.

RC: 18. Line 99: Is the "Eulerian curvilinear grid" the coarse-grained grid or the finer grid?

AR: Thank you for seeing this lack of detail, we modified the sentence to assess that the interpolation is done on the fine grid and the fields are then coarse-grained. The sentence became: "Interpolated onto the native Eulerian curvilinear grid with nearest neighbors, forcings at time t , $t+6h$ and $t+12h$ are then coarse-grained, normalized and added as predictors to the input of the neural network".

RC: 19. Lines 105 and 190: Please specify units for the observation-error variance.

AR: The observation-error variance is defined in the normalized space, so it has no units. We stressed on this in line 105: "Through all experiments, $\sigma_b^2 = 0.4^2$ (unitless)."

RC: 20. Line 108, the first part of Equation 2b: As far as I understand, the tilde quantities are normalised quantities, so shouldn't the equation be $\tilde{x}_{t,\text{obs}}^{\text{LN}} = (\mathbf{x}_t \exp(\dots) - \mu)/\sigma$

AR: Thank you for seeing this mistake in the equation this is indeed how the equation is computed.

RC: 21. Line 109: The notation \min_{SIT} is confusing. With SIT in the subscript, I thought, at first sight, that it meant the minimum of something over SIT, with the argument (the "something") lacking.

AR: Thank you, we clarified its definition: "Furthermore, a variant to log-normal noise is introduced in Eq.(2c) by adding a fraction of the sea-ice thickness and incorporating clipping, based on SIT_{\min} the corresponding 0m thickness in the normalized space."

RC: 22. The sentence spanning over lines 110 – 111: There are probably too many unnecessary commas that hinder my ability to understand this sentence.

AR: Thank you for your comment, we improved the sentence. "The Gaussian observation noise as defined in Eq.(), is an idealized case tailored to the common assumptions of 4D-Var, to test an adaptive inflation scheme which will be defined later"

RC: 23. Line 113: You say that a variant to log-normal noise is introduced in Equation 2c, but where is the log in the equation?

AR: Thank you, we meant a variant to the log normal, in the sense that, as for the log normal noise, it applies noise of higher magnitude where there is more ice. The similarity between the two noises can be seen in Fig.1. We modified the sentence: "This approach ensures that the observations remain confined within the physical bounds of sea ice, unlike Gaussian noise, but similarly to log-normal noise."

RC: 24. Section 2.2.2, opening paragraph: A short elaboration about the complementarity of CryoSat-2 and SMOS observations would be desirable here.

AR: Thank you for your remark. The sentence "The retrievals merge observations from CryoSat-2 (Kurtz and Harbeck, 2017), known for its observations of thick and perennial sea ice, and from SMOS (Tian-Kunze et al.,

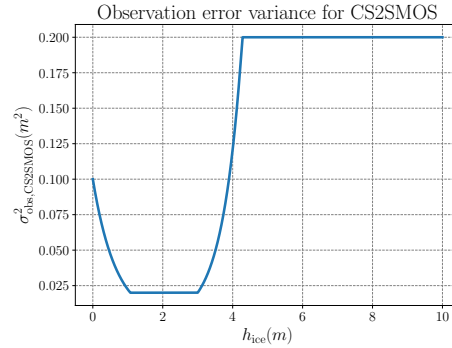


Figure 1: Representation of the function defined in Eq. 3.

2014), used to infer the thickness of thin ice.” indicates their complementarity. We refer to Ricker et al. (2017) for more specific details. We changed the sentence to ”The retrievals merge observations from CryoSat-2 (Kurtz and Harbeck, 2017), known for its accurate observations of thick and perennial sea ice, and from SMOS (Tian-Kunze et al., 2014), used to infer the thickness of thin ice.”

RC: 25. Line 125: What is “Kriging”?

AR: Kriging is a widely accepted term for optimal interpolation methods in space.

RC: 26. Line 132: Can you provide a graph of the function described in Equation 3, for easier visualisation? Please also define the units used in the coefficients for this equation.

AR: Please find in Fig. 1 of the present document the required graph. The coefficients dimensions were added in the paragraph: ”Based on the diagnostics of Desroziers et al. (2005), Xie et al. (2018) proposed an empirical formula for the observation error variance $\sigma_{obs,CS2SMOS}^2$, as an increasing function of ice thickness h_{ice} , with the coefficient 0.2, 0.02, 0.1 in m^2 , 3 in m and 1.5 without unit,”

RC: 27. Lines 134 – 136: This is not a problem, because you do this in a pre-processing step. The assimilated “observations” are actually retrievals.

AR: Thank you for your remark, we believe it is worth to mention it. We added a comment: ”In a preprocessing step, real observations are interpolated onto the model space, making them retrievals.”

RC: 28. Lines 136 – 137: Do you thin the observations or create super-observations? If not, how confident are you that the observation errors are negligibly correlated?

AR: Thank you for your questions. We changed the sentence to 'This is feasible because the observations are at a higher resolution in their native grid, and then coarse-grained by a factor 2.' Note that in any case, the observation error variance is computed after the interpolation on the coarse-grained grid. Thus we are not creating 'super-observations'. The notion of thinning would be more relevant here.

RC: 29. Lines 144 – 147: How do you handle the non-differentiability of the ReLU function at the kink (in your case, at $\min()$ *)?

AR: In tensorflow the ReLU function gradient value is set to 0 at the break point.

RC: 30. Line 150: Is $N_x \times N_y$ the same as the 128×128 grid you mentioned earlier? Are the masked cells (due to land) excluded?

AR: Yes thank you for your remark, we added this: "The main part of the loss function is defined by a pixel-wise mean-squared error (MSE) on all $N_x \times N_y = 128 \times 128$ grid-cells, multiplied by the land-sea mask".

RC: 31. Line 159: Why is the clipping in Equation 4b omitted during the training?

AR: The idea is to first focus on learning the dynamical changes represented by f_θ . Since we are learning the difference between two SIT fields, the output, including errors, the future SIT state may become non-positive. However, we have observed that learning the differences instead of the full next state improves the emulator's forecasting ability. Once this emulator effectively captures SIT evolution, we incorporate the positivity constraint of the SIT variable through fine-tuning.

RC: 32. Line 160: Is there a reason for using such a big weight at the initial training stage, but not at the second stage?

AR: As mentioned in Lines 161-162: "Here, we select $\lambda = 10$ as penalty weight, striking a good balance between $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ which have a different magnitude during second training." The magnitude of the losses $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ are different between the first and the second training. This is simply due to the different minimization objectives. The "big" weight during first training correspond to a ratio between $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ of 100 : 1. With a weighting of 10 during the second training, we obtained a comparable ratio to the first training (cf Fig. A1).

RC: 33. Section 4.1: It would be nice to mention somewhere that this is full 4D-Var (if I understood correctly) instead of the incremental 4D-Var that is commonly used in NWP centres. It took me (as a reviewer) a while to figure out that the 4D-Var experiments here are conducted in full-field space instead of incremental space.

AR: Thank you for your remark, we stressed this point with the sentence: 'Note that we are focusing on full 4D-Var rather than incremental 4D-Var.'

RC: **34. Lines 169 – 170 and 505: Presumably multiple observations valid at the same time are assimilated. How many observations are used at each observation time? The use of “ k^{th} - observation” should be avoided if you mean the vector of observations at the k^{th} - observation time. Similarly, in line 505, it should be the number of observing times per cycle instead of the number of observations per cycle.**

AR: Thank you for your remark. Observations are assimilated on every valid grid-cells (8871 obs per state). We changed by specifying 'observing times' every time needed. We will change the sentence in L169: " \tilde{y}_k represents the observation at time k "

RC: **35. Line 170 and beyond: If everything from this point onwards is in normalised space, then it would be desirable to drop the tilde notation.**

AR: We feel that this would confuse even more the reader. Especially since all the metrics in the results are computed back into the physical space, as outlined by their definitions in Sec. 5.1. We did not remove the tilde.

RC: **36. Line 171: Without having read the following sub-sections, a reader would naturally wonder what the specifications for B are. It may be useful to mention here that such details are given in the following sub-sections.**

AR: Thank you for your remark, we added this info: " B is the background error covariance matrix and its specific formulations will be provided below."

RC: **37. Lines 173 – 174: Do you mean that the background for the first cycle is the neXtSIM “truth”? Idealised DA experiments should begin with a background that is the truth perturbed by random statistics of the B matrix, rather than the truth itself.**

AR: It is not the truth as it is a snapshot from a previous year (2016). The RMSE between the first guess and the truth is equal to 0.37 m.

RC: **38. Line 174: You start the experiments before the training period was over. Are there any implications?**

AR: The experiments starts on 2017-01-01, which is after the training period. Observations and forcings are taken at the correct dates (so outside the training dataset). Only the first guess of the first cycle comes from the training dataset. Therefore, there is no risk of data leakage in our experiment.

RC: 39. Line 176: What is the “that” referred to in “After that”? Also, do the “both cases” refer to 4D-Var-diag and 4D-Var-EOF?

AR: Thank you for seeing the error. We changed the sentence to: “The observation operator \mathcal{H} is simply defined as a diagonal matrix \mathbf{H} .”

RC: 40. Lines 176 – 177: Is the \mathbf{B} matrix simply a matrix of ones and zeros, a sub-selection of rows of the identity matrix (depending on your observation coverage)? \mathbf{B} won’t be a diagonal (square) matrix unless you have exactly the same number of observations as the number of elements in your state vector.

AR: Yes, exactly we have the same number of observations as the number of elements in our state vector.

RC: 41. Line 178: What is N_{cycle}

AR: From the sentence ‘In this study, the 4D-Var is cycled across N_{cycle} cycles.’ we get that N_{cycle} is the number of 4D-Var cycles. Its numerical value change between the twin experiment and the real observations cases. We recalled the variable in each subsection: at L239: “The 4D-Var is run on a single trajectory starting the January 1st 2017, for $N_{\text{cycle}} = 45$ cycles, until December 21th, 2018. ” and at L266: “In this section, instead of assimilating simulated observations, we assimilate CS2SMOS retrievals daily (every two iterations of the emulator) within an 8-day window for $N_{\text{cycle}} = 20$ cycles.”

RC: 42. Line 181: I wouldn’t regard this as two “types” of 4D-Var. They only differ by the choice of the \mathbf{B} matrix effectively. For the 4D-Var-EOF case, according to your description in lines 201 – 202, it is equivalent to having $\mathbf{B} = \varphi_m \varphi_m^\top$

AR: Thank you for your remark. The difference is indeed on the \mathbf{B} expression. But in practice we also change the variable on which the minimization is conducted. Thus we believe that the wording ‘two types’ is acceptable.

RC: 43. Line 181: It is not appropriate to refer to \mathbf{B} as the “background” matrix. It is a covariance matrix of background errors, not of the background itself. Similarly, you are inflating the background errors, but not the background, in lines 191 and 340, 507 and 508.

AR: Thank you for your remark, we corrected those instances: line 181: “Two types of 4D-Var are evaluated. 4D-Var-diag, which correspond to the use of a diagonal matrix as covariance matrix of background errors \mathbf{B} ”, line 191: “The coefficient λ_{inf} is a background errors multiplicative inflation term, which is set to 1 when no inflation is used and is further described”, line 340: “ The adaptive background error inflation can be easily implemented in twin experiment scenarios, under a Gaussian noise simulation.”,

line 507: "We adopt an adaptive multiplicative inflation scheme on the background error term".

RC: 44. Lines 192 – 193: Generally speaking, inflating the observation-error covariances is not equivalent to inflating background-error covariances. This is because the \mathbf{B} matrix in 4DVar is implicitly propagated in time by the linearised model \mathbf{M} (in order words, \mathbf{B} becomes $\mathbf{M}\mathbf{B}\mathbf{M}^\top$). I would like to see how the covariances or correlations look like when \mathbf{B} is implicitly propagated by the linearised model \mathbf{M} .

AR: We changed the wording of the sentence to be more precise: "Note that the observation error covariance matrices \mathbf{R}_k are not inflated because, with diagonal matrices, this is equivalent to background error covariance matrix inflation, as far as the cost function minimization is concerned". Indeed, in case of diagonal matrix, for the cost function, inflating one term is equivalent to deflating the other term.

RC: 45. Lines 199 – 200: When you say "The projection onto the EOFs enables access to cross covariances", I imagine you say this as an advantage compared to the 4D-Var-diag case where there is no cross-covariance. However, the diagonal \mathbf{B} case is quite extreme and impractical, so I wouldn't say this as a feature of using EOFs. You also suggest that the projection onto EOFs improves numerical conditioning. Yet, since the \mathbf{B} matrix for the 4D-Var-diag case is a scalar multiple of the identity matrix, its condition number is 1 and therefore the numerical conditioning cannot be improved beyond that.

AR: Thank you for your remark. Yes, the EOFs variant of the 4D-Var accounts for background cross-covariances as opposed to the 4D-Var-diag case. Expressing the cost function as a function of the EOFs coefficients improves the conditioning of the optimisation problem (providing a lower bound for the Hessian), not the background error covariance matrix per se, and compared to the full 4D-Var with a cost function written in state space. We have modified this sentence to "The projection onto the EOFs enables access to cross-covariances compared to the 4D-Var-diag case and betters the numerical conditioning of \mathcal{J} compared to the state space 4D-Var without approximation of the background error covariance matrix, thereby improving the minimization of the cost function."

RC: 46. Line 208: It is better to express the second term of the cost function in terms of \mathbf{w}_0 .

AR: Thank you for your remark, we agree with you and we changed it accord-

ingly:

$$\mathcal{J}(\mathbf{w}_0) = \frac{1}{2\lambda_{\text{inf}}^2} \|\mathbf{w}_0 - \mathbf{w}_0^{\text{b}}\|^2 + \frac{1}{2} \sum_{k=1}^K \left\| \tilde{\mathbf{y}}_k - \mathbf{H}_k g_{\theta}^{k \times N_f}(\bar{\mathbf{x}} + \varphi_m \mathbf{w}_0, \mathbf{F}_{k \times N_f \rightarrow 0}) \right\|_{\mathbf{R}_k^{-1}}^2. \quad (1)$$

RC: 47. Line 210: I think you should state the choice of the truncation index here. This is important information that shouldn't be relegated to an appendix.

AR: We emphasized on the chosen value for the truncation index, and refer to the appendix for further discussion: "The value of the truncation index is set to $m = 7000$ for all experiments and is further discussed in the Appendix() while further details on the optimization are given in Appendix()."

RC: 48. Line 222: Is there a reason to average the RMSE over all grid points but keep it as a function of time? Unless there is a good reason to look at the results for individual assimilation cycles, I think it would be more interesting to evaluate the RMSE by taking the mean over assimilation cycles yet presenting the results as a function of spatial location (i.e. map plots).

AR: We understand the potential interest of adding map plots to see where the errors are localized. The evaluation of the RMSE as a function of time allows a more simple evaluation. As we see in Fig.5, there are important changes in the RMSE during the year that would not be visible with map plots, which justify our use of this metric.

RC: 49. Lines 228 – 232: IIEE is standard terminology. If you use a modified definition, it is better to give an alternative name to it, to avoid being misleading. Do you have any evidence demonstrating that an SIT threshold of 0.1 m is roughly equivalent to an SIC threshold of 0.15? In any case, it would be good to show results after weighing them by grid cell areas, especially when the grid cell areas are not uniform.

AR: We originally discussed this choice in Durand et al. (2024), cf Fig B1 (in this paper, our modified IIEE metric was called acc_{SIE} . We can also show the scatter plot of neXtSIM SIC vs SIT in Fig. 2 of the present document. In this scatterplot, we can see that the value of 10 cm corresponds to the transition from SIC close to 0 to SIC close to 1. We linked the original justification by changing the sentences: "A grid cell is considered to be covered by sea ice if the thickness exceeds 0.1 m (cf (Durand et al., 2024)-B1 for the threshold justification)" We presented results in the new version with the IIEE weighted by grid-cells area and we used the terminology IIEE_{SIT} .

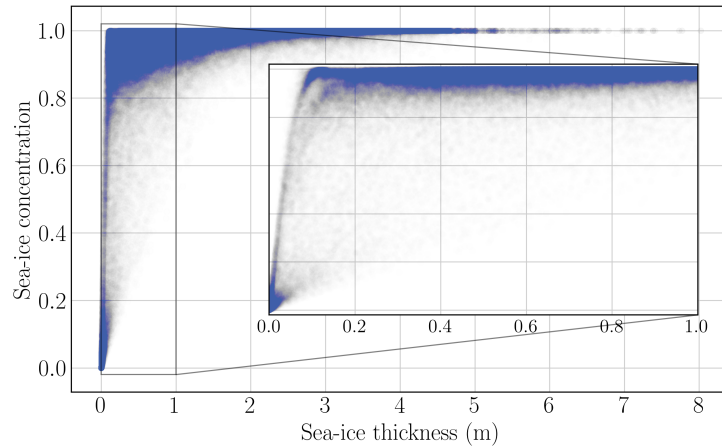


Figure 2: Scatter plot of the sea-ice concentration depending on sea-ice thickness in January 2017. The zoom provides more details information of the transition between 0 and 1m of sea-ice thickness.

RC: 50. Line 235: What do you mean by “the same past field”?

AR: Thank you for your remark. We meant that all experiments are initialized for the first SIT field of the first cycle with a field from neXtSIM (2016-01-01). We changed the sentence to make it clearer: "In all twin experiments, we initialize the first data assimilation cycle with a past neXtSIM SIT field from the 1st of January 2016."

RC: 51. Line 237: For the cycle that spans from late 2017 to early 2018, is the result discarded?

AR: Yes, the first 'test cycle' start on 2018-01-04.

RC: 52. Line 238: “45 cycles” means 720 days, so the last cycle should end on 21 December 2018 (end of the day), not 22 December.

AR: Thank you for your remark, this is an error that will be corrected, the sentence is now "The 4D-Var is run on a single trajectory starting the January 1st 2017, for 45 cycles, until December 21th, 2018."

RC: 53. Line 241 and caption of Table 1: Why would you like to assimilate perfect observations (taken from the truth)? If you do that, please define “the RMSEs between neXtSIM and the perturbed observations” mathematically, as you do with the other RMSE definitions.

AR: Thank you for your remark. The results presented here serve primarily as a proof of concept. Initially, we were unsure whether assimilating observations into our emulator using 4D-Var minimization would be feasible. The use of perfect observations, while not realistic in practice, was intended to simplify the problem. We defined the Observation RMSE in the caption of the Table 1: "They correspond to the RMSE of each observation to its associated non-perturbed field."

RC: 54. Line 242: If the results are comparable and there is no discussion about the advantages and disadvantages of different noise types, then perhaps you don't have to show the results for all 3 types of perturbations.

AR: Thank you for your remark, we do discuss the results on the noise first, and afterwards, we only show results with the cond-clipped noise (which is not the most "natural" noise definition. Yet, especially for the inflation analysis in the appendix, it was necessary to introduce the Gaussian noise. We will keep the results display as it is.

RC: 55. Line 242: "type" should be in the plural.

AR: Thank you, we corrected this.

RC: 56. Line 245: I think the improvement is more likely caused by the realism of the B matrix instead of the preconditioning and off-diagonal terms.

AR: Thank you for your remark. We modified the sentence: "This improvement can be attributed to the non-diagonal terms in the **B** matrix."

RC: 57. Figures 3 and 7: Some characters are missing in the legend of the graphs.

AR: Thank you for your remark, it is not the case in our overleaf pdf, we will ensure that the labels are not corrupted in the next version.

RC: 58. Figure 3: I am a bit surprised by the small difference between the grey and purple curves. Could you extend this graph to show days 0 to 15 as well? I would expect the difference to be larger at the beginning of the assimilation window, where the difference in the B matrix is most significant. In a sense, as the assimilation window goes on, the difference might diminish by the fact that it is the same linearised model M that propagates B (cf. Minor Comment 44).

AR: Yes, we originally added the first 16 days, and we show you the figure below in Fig. 3 of the present document. Yet, as this period corresponds to the analysis in the DAW, we decided to remove it to make the figure easier to understand. As you can see, the gap between the two types of 4D-Var is decreasing as the lead time increases.

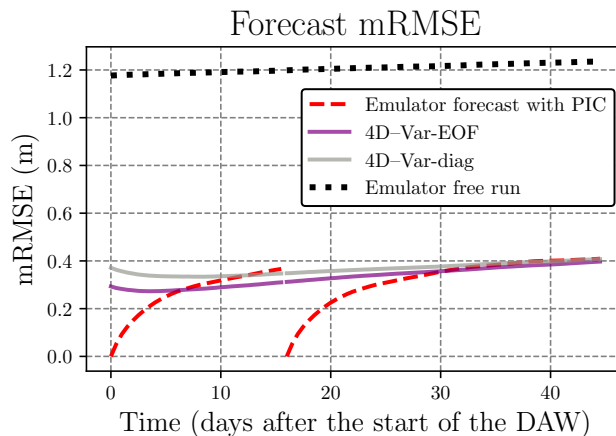


Figure 3: Cycle-averaged mRMSE of $\tilde{\mathbf{x}}_f$ over 2018 during the analysis and forecast stages. The dashed red line represents the free run of the emulator started at the beginning (time 0) and at the end (time 16) of each DAW (with perfect initial conditions, PIC). The mRMSE of the 4D-Var-EOF forecast is shown in purple, and the 4D-Var-diag forecast is shown using in grey. Cond-clipped noise is used in both cases. The black dotted line corresponds to the emulator free run, initialized on January 1st, 2017.

RC: 59. Figure 3: The horizontal axis here is not an absolute date, but in terms of days after the start of the DAW. How do you compute the black dashed line in this case? Why is it not flat, but instead changes slightly with increasing lead time?

AR: The free run is initialized at the beginning of the first cycle. Then, for each cycle, we take the free run RMSE at the corresponding absolute time, resulting in an increasing function. These RMSE values are then averaged (mRMSE) in the same way as for the other lines. Consequently, we obtain a slightly positive function, consistent with the mRMSE definition.

RC: 60. Line 250, regarding the words “we observe a slight improvement of 1.1 cm”: Do you mean towards the end of the forecast when the red dashed line in Figure 3 is above the purple line? 1.1 cm seems to be minimal. Do you know whether the result is statistically significant? If so, how would you explain it (the 4D-Var-EOF case being better than the emulator forecast from perfect initial conditions)?

AR: Thank you for your remark. This is the improvement we meant. Please see in Fig. 4 of the present document the same figure with additional display of standard deviation (not shown in the paper for the figure clarity) and zoomed in on the last 15 days of the forecast. We removed the free run line

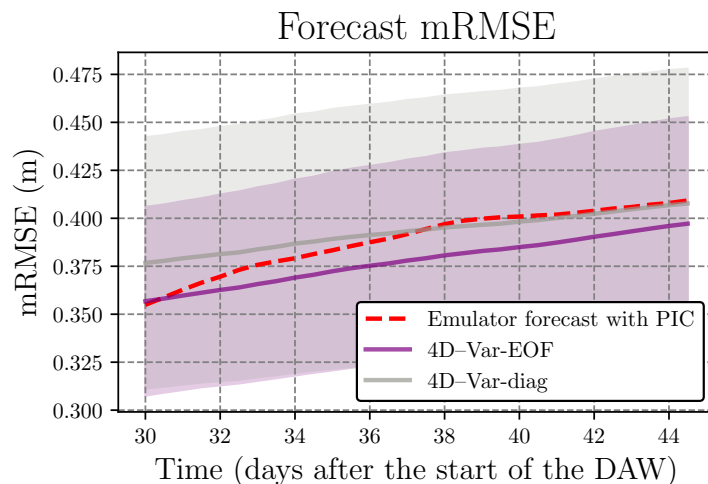


Figure 4: Cycle-averaged mRMSE of \tilde{x}_f over 2018 during the last 15 days forecast stage. The dashed red line represents the free run of the emulator started at the end of each DAW (with perfect initial conditions, PIC). The mRMSE of the 4D-Var-EOF forecast is shown in purple, and the 4D-Var-diag forecast is shown using in grey. Cond-clipped noise is used in both cases.

to focus on the part in question. The improvement is small but visible. We changed the sentence to "Additionally, when comparing the 4D-Var-EOF forecast to the emulator forecast (initialized with perfect conditions at the end of each DAW, red dashed curve), we observe comparable RMSE, showing the stability of the forecast produced by the analysis."

RC: 61. Lines 252 – 253: Do you know why the analysis is noisier than the background? Is it because the forecast (using the emulator) smooths out features?

AR: Thank you for your remark. The noise in the analysis comes from the observation noisiness and the background is indeed smoother because of the emulator. We revised the sentence: "The first guess field is smoother because of the emulator, whereas the analysis is actually noisier than the truth, because of the observations noisiness."

RC: 62. Line 255: Is the negative bias a known issue of the surrogate model, or is it only an issue at this time of the year?

AR: Thank you for your remark. As seen in Fig. 5, this negative bias is an issue at this time of the year. The sentence "In the other places, we observe a positive correction, which is consistent with the negative bias ($\simeq -0.015$ m) of the surrogate model at the time of the depicted cycle." reflects this.

RC: 63. Figure 4: Is this for 4D-Var-diag or 4D-Var-EOF? If it is 4D-Var-diag, the only reason that you are not getting point-based increments is the implicit propagation of B by the linearised model M (cf. Minor Comment 44).

AR: Thank you for your remark. It is the 4D-Var-EOF. We corrected the caption accordingly: 'Fields of the SIT in the 10th cycle of the DA, corresponding to 2017-06-10, are shown in the 4D-Var-EOF case.'

RC: 64. Caption of Figure 4: a. Please state the full 16-day range of the assimilation window instead of just "2017- 06-10". b. It is not clear what the valid date / time of the "forecast" and "analysis" is. c. Why do you choose to show the results of this cycle? You said earlier that you treat 2017 as a spin-up year.

AR: a- We focus on the first day of the analysis since we look at the output of the minimization - we changed the caption accordingly: 'Fields of the SIT in the beginning of the 10th cycle of the DA, corresponding to 2017-06-10, are shown in the 4D-Var-EOF case.'

b- cf above

c- Indeed we picked a random cycle, we selected one in 2018 for the revised manuscript on 2018-02-05, cf Fig. 5 of the present document.

RC: 65. Caption of Figure 5: How do you launch a new forecast every 6 hours when the assimilation cycle is 16 days long

AR: The emulator forecast is initialized every time we have a sample in the neXtSIM outputs (every 6 hours in 2018) and run for 16 days, which is how we obtain this bias error.

RC: 66. Line 256: There are exceptions to "the initial analysis RMSE [being] lower than the first guess RMSE" as indicated in the top panel of Figure 5. For example, in late February 2018, the analysis RMSE is larger than the background RMSE. Do you have any idea why that is the case?

AR: Thank you for your remark, as mentioned in the text "As seen in Fig.5, at the start of each cycle, the initial analysis RMSE is generally lower than the first guess RMSE", there are exceptions. In the specific case on late February 2018 (cycle 26), as seen in Fig. 6 of the present document. Large differences between the truth and the analysis are located near the Canadian Archipelago, leading to a higher RMSE compared to the first-guess RMSE, where such differences do not occur. We hypothesize that, in order to minimize the RMSE throughout the DAW, the 4D-Var removes sea ice in this region. This is consistent with the emulator's positive bias during this period. We added a sentence after L.256 to raise this point, but we did not add the Fig. 6 of the present document: "In the specific case on late February 2018 (cycle 26), large differences between the truth and the

4D-Var, 25th cycle

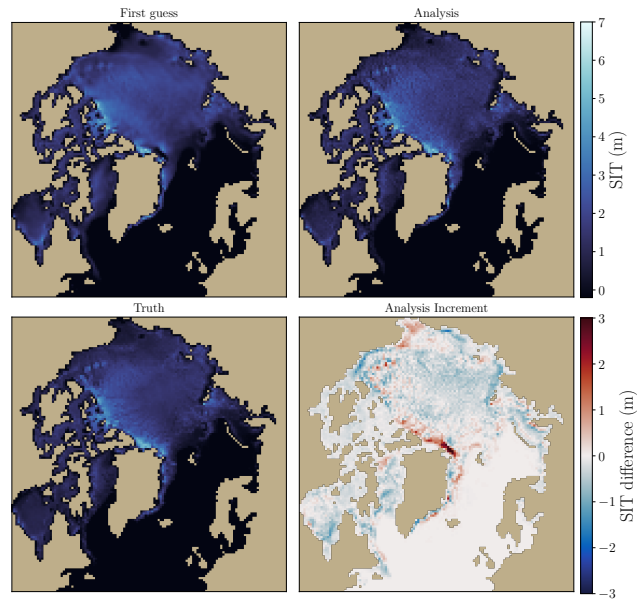


Figure 5: Fields of the SIT at the beginning of the 25th cycle of the DA, corresponding to 2018-02-05, are shown in the 4D-Var-EOF case. The upper left panel represents the first guess, which is the output of the forecast from the previous minimization. The upper right panel corresponds to the analysis of the 25th cycle. For comparison, the associated neXtSIM field, considered as the truth, is displayed in the lower left panel. Note that these three fields share the same colormap and scale. The lower right panel shows the analysis increment, which represents the analysis minus first guess.

4D-Var, 26th cycle

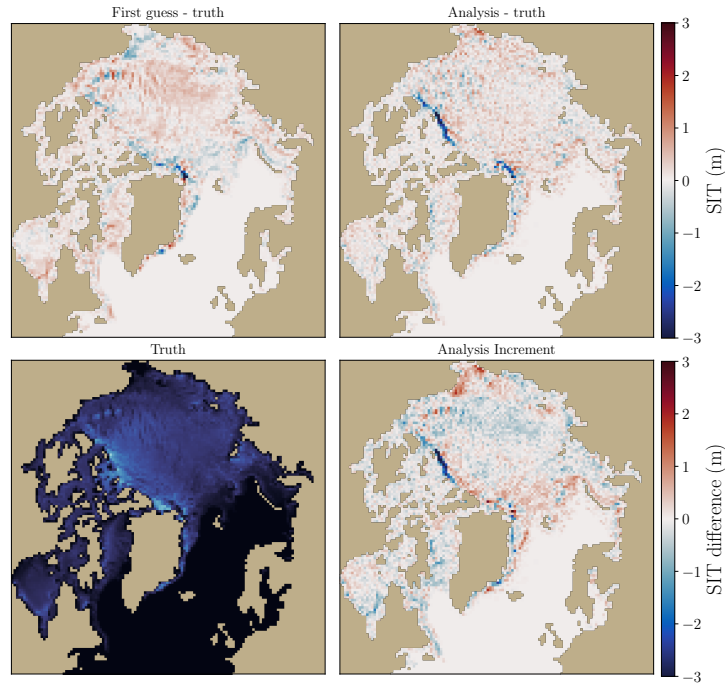


Figure 6: Fields of the SIT at the beginning of the 26th cycle of the DA, corresponding to 2018-02-21, are shown in the 4D-Var-EOF case. The upper left panel represents the first guess minus the truth. The upper right panel corresponds to the analysis of the 26th cycle minus the truth. For comparison, the associated neXtSIM field, considered as the truth, is displayed in the lower left panel. The lower right panel shows the analysis increment, which represents the analysis minus first guess.

analysis are located near the Canadian Archipelago, leading to a higher RMSE compared to the first-guess RMSE, where such differences do not occur. We hypothesize that, in order to minimize the RMSE throughout the DAW, the 4D-Var removes sea ice in this region. This is consistent with the emulator’s positive bias during this period.”

RC: 67. Lines 258 – 259: I would say the peak is in June and not May, and I think it is difficult to conclude that the RMSE is rising again in December just by looking at the top panel of Figure 5.

AR: We corrected the sentence for July. Regarding the raise in December, it was clearly visible in the 2017-2018 full plot for the end of 2017. Since those results are discarded, we removed the end of the sentence. It became: ”The

analysis from Fig.5 reveals a strong seasonality in results, with RMSE peaking in July. The peak aligns with significant changes in sea-ice extent at this period, with an important decrease due to the raise of temperature in the Arctic.”

RC: 68. Line 260: Without further evidence, I would be reluctant to conclude “4D-Var struggles more with dynamic shifts” only based on the coincidence between the trends in the top panel of Figure 5 and the seasonal changes in sea-ice extent. As you say in the next sentence, it could have something to do with the bias of the emulator.

AR: Thank you for your remark. The emulator also struggle because of the dynamical changes. Hence, we kept the sentence but we mitigated it even more: ”This suggests that 4D-Var possibly struggles more with dynamic shifts”.

RC: 69. Line 266: Do you use (northern hemisphere) winter because CryoSat-2 data are unavailable in the summer?

AR: Thank you for your remark, we changed the sentence: ”We use the observations for the winter 2020-2021, as there are no CS2SMOS retrievals during summer” to make this point clear.

RC: 70. Lines 266 – 268: Would it be possible to randomly select a subset of CS2SMOS observations and not assimilate them, but rather use them for verification? This would give better confidence in the interpretation of results.

AR: In any case, the observations are strongly correlated in time. If we were to remove enough observations to reduce this correlation between assimilation and verification, we would not have enough left to assimilate. By using observations after a 9-day forecast for comparison with neXtSIM-F and subsequent interpretations, we move beyond the temporal averaging window of CS2SMOS. We believe this is sufficient to trust the results.

RC: 71. Lines 268 – 269: The date 10 October 2016 is a long way before the verification period (winter 2020 – 2021). Is there a reason to start so early? Also, this date is inconsistent with the date given in the caption of Figures 6 and 7.

AR: Thank you for seeing this, this date is inconsistent. We use for the first guess of the first state the actual CS2SMOS field. We modified the sentences and caption accordingly ”To initialize the cycling of the data assimilation, we start with the CS2SMOS retrieval from the date of the first assimilation cycle.” However, the caption in Fig 6 and 7 are correct as it corresponds to the emulator start.

RC: 72. Line 270: I note that the inflation of the B matrix is not used in both the idealised experiments and CS2SMOS experiments. In that case, does it need to be mentioned in the paper (whether in an appendix or not)?

AR: Thank you for your remark. We believe it is a point worth mentioning in the Appendix.E

RC: 73. Lines 271 – 273: I wonder how independent neXtSIM-F is compared to your neXtSIM based emulator, given that both are related to neXtSIM. Could you elaborate on that?

AR: neXtSIM-F runs on a different grid and is forced with TOPAZ ocean forcings. The emulator operates at a much coarser resolution and was trained on a simulation where neXtSIM was coupled to NEMO. Since we always compare our emulator to neXtSIM, it was natural to also compare it with its associated forecasting product. However, the significant resolution difference (3 km vs. 50 km) makes them independent enough.

RC: 74. Lines 274 – 275, “NeXtSIM-F assimilates... with a simple nudging”: This is not clear. Are you assimilating or nudging CS2SMOS observations?

AR: Thank you, we modified the sentence: ”NeXtSIM-F nudges CS2SMOS sea-ice thickness observations weekly.”

RC: 75. Line 275: Is the 9-day forecast from the start or end of the assimilation window?

AR: The 9 day forecast starts at the end of the assimilation. We simply use the neXtSIM-F operational results. We download the files which started at the corresponding time (end of each DAW) and then use the 9-th day forecast file to compute the RMSE with CS2SMOS. We added some details at the end of the sentence: ”It produces a 9-day forecast, which we systematically use for numerical comparison with our data assimilation scheme, starting the neXtSIM-F forecast at the end of each DAW.”

RC: 76. Lines 276 – 277: If I understand correctly, you are using neXtSIM-F only as a reference (or “independent”) gridded forecast dataset for verification purposes. Why do you have to compare neXtSIM-F against CS2SMOS observations?

AR: We do this to compare how well both forecast systems perform against CS2SMOS.

RC: 77. Line 278: Could you elaborate on what “fairness” you refer to?

AR: ERA5 is a reanalysis product that includes observations, making it 'un-fair' to use in an operational setup where the goal is to predict the future. Therefore, as done in neXtSIM-F, we use atmospheric forecasts for the additional atmospheric forcings provided to the emulator during the forecast period.

RC: 78. Line 279: “that are run” should be “that is run”.

AR: Thank you we corrected it.

RC: 79. Line 279: Is there a reason to force the model with HRES instead of ERA5?

AR: Please refer to the point 77. We added a sentence after L279 to make the point clearer: "In contrast, ERA5 is a reanalysis product that assimilates observations, making it unsuitable for an operational forecasting setup where the aim is to predict the future. Therefore, following the approach used in neXtSIM-F, we rely on atmospheric forecasts as forcings of the emulator during the forecast period"

RC: 80. Lines 283 – 284: How difficult is it to find an initialisation field in 2020 for your winter 2020 – 2021 experiment? Could we get round this problem?

AR: Thank you for your remark. We made a mistake in the text. In the end we decided to initialize the field with a CS2SMOS observation from 2020.

RC: 81. Line 286: “In average” should be “On average”.

AR: Thank you, we corrected this.

RC: 82. Lines 287 – 288: Do you know if there is anything that drives the trend in the difference between the experiment’s 9-day forecast RMSE and the neXtSIM-F 9-day forecast RMSE?

AR: Thank you for your remark. We hypothesize that, since SIT follows a log-normal rather than a Gaussian distribution, its standard deviation increases with the mean during winter. This effect could be amplified by the presence of leads in neXtSIM-F and their absence in CS2SMOS. In contrast, our emulator provides smoother forecasts, which may explain why it does not exhibit a similar RMSE increase.

RC: 83. Lines 288 – 289: Do you have evidence demonstrating the claim? If double penalty is the issue, it might be more suitable to use the Fractions Skill Score (see e.g. DOI: 10.1175/2007MWR2123.1).

AR: One can see the local dynamics in the neXtSIM-F field. Double penalty might be a strong claim and we will revise the sentence accordingly. Please see in Fig. 7 here an example of the differences between CS2SMOS, neXtSIM-F and our forecast. We do not see 'real' double penalty effect

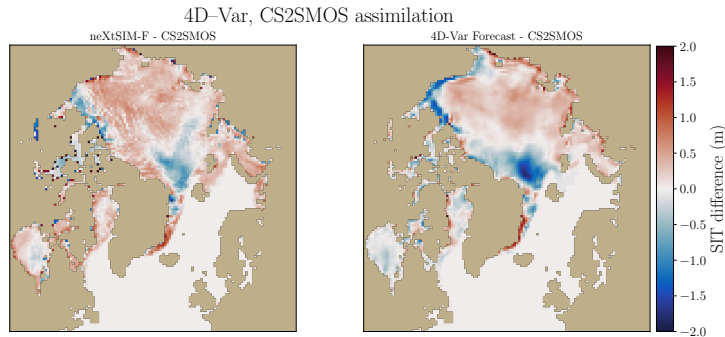


Figure 7: Left: Difference between neXtSIM-F SIT and CS2SMOS SIT on 29-01-2021. Right: Difference between 4D-Var 9 day forecast SIT and CS2SMOS SIT on 29-01-2021.

since the leads are not misplaced with CS2SMOS but rather inexistent. We changed the sentence to: "The RMSE metric penalizes the higher level of details of the neXtSIM model more than the emulator that smooths gradually with time." Let us note that in Fig. 7 here we also see on some pixels big local errors where neXtSIM-F places sea ice when CS2SMOS does not observe sea ice (typically in Canadian Archipelago). Please also see in Fig. 8 the results for the Fraction Skill Score metric you proposed. We relied on the implementation of the metric from the pysteps package (Pulkkinen et al., 2019), and we chose a threshold of 0.2 m and a scale of 3 pixels.

RC: 84. Lines 290 – 291: More evidence is needed to support the claim that the DA system is “less efficient during periods of strong dynamic change”. Perhaps you could consider running a multi-year experiment.

AR: Thank you for your remark. We would need multi-year observations to do this in the real observations case. We show this in the twin experiment case. We displayed in Fig. 9 of the present document the results for a 2-year twin experiment. We observe that the increase in RMSE occurs during the same months each year—when sea ice melts in June and refreezes in November. We added this figure to the Appendix to emphasize this point.

RC: 85. Lines 291 – 293: The sentence is not clear. Which are the “dynamic” periods you refer to?

AR: Thank you for your remark. We were mentioning the refreezing period in October-December, as opposed to the colder period coming afterwards. We rephrased this part: "Conversely, after the refreezing period, it faces fewer difficulties in predicting the optimal state."

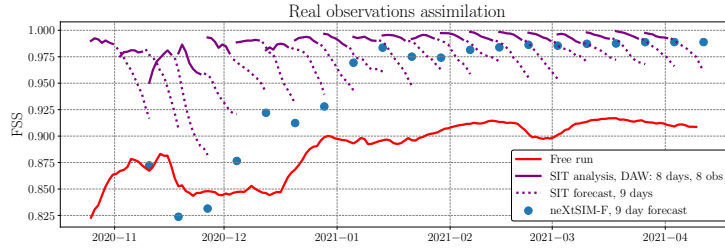


Figure 8: FSS results for CS2SMOS assimilation across several DAWs throughout the full CS2SMOS observation period in 2020-2021 are shown. The red line represents the free run of the emulator through all cycles, initialized with a SIT field from October 2018. The solid purple line corresponds to the analysis of the 4D-Var over the DAW, while the associated dotted purple lines represent the additional forecasts using ECMWF atmospheric forecasts for 9 days. The FSS from neXtSIM-F, corresponding to 9 day forecast, are displayed as blue dots and should be compared with the end of each corresponding dashed line. All scores are computed with CS2SMOS considered as the truth.

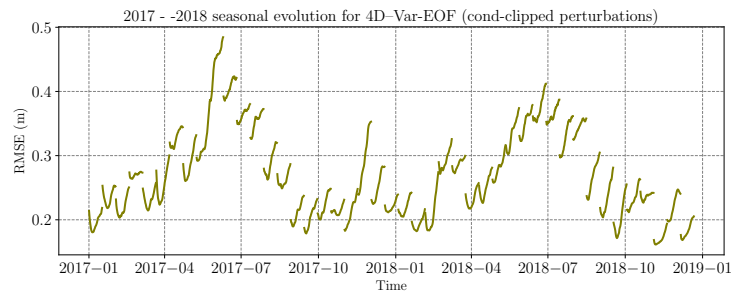


Figure 9: Time evolution of the 4D-Var-EOF analysis RMSE (inside the DAW) in 2017 and 2018, with cond-clipped simulated noise.

RC: 86. Line 295: What “trajectory” do you mean?

AR: We apologize, we meant “free run”. We corrected the sentence accordingly: “The bias error of the emulator free run is significantly reduced.”

RC: 87. Line 298, regarding the words “the bias systematically decreases”: During the forecast stage (dashed purple lines in the top panel of Figure 7), the value of the bias indeed systematically decreases, but depending on whether the bias is above or below zero, it could mean either a worsening or an improvement. The wording needs to be more careful here.

AR: Thank you for your remark. We changed the sentence to “The absolute value of the bias of the emulator free run is generally reduced, with the exception of December 2020, where the free run bias is smaller than the forecast bias.”

RC: 88. Lines 298 – 299: I am not sure what you mean by “the emulator considers the analysis to increase the amount of sea ice”.

AR: Thank you for your remark. What we meant is that since the emulator tends to reduce the amount of sea ice (systematic decreasing bias), the analysis systematically has more sea ice than the observations (positive bias at the start of the analysis) to counter this decrease. We changed the sentence to: “The biases at the beginning of the DAWs are systematically positive and then generally decrease. Since the emulator tends to reduce the amount of sea ice, as indicated by its negative bias, the analysis overestimates sea ice to compensate for this loss.”

RC: 89. Lines 299 – 300: Is there any evidence demonstrating your claim, namely that neXtSIM having a different SIT distribution from CS2SMOS could explain the phenomenon?

AR: The emulator is trained on neXtSIM data, meaning its output distribution will match the neXtSIM SIT distribution. If we attempt to match CS2SMOS with the neXtSIM SIT field (see Fig. 2), it is expected to struggle and tend to revert to the neXtSIM distribution. The strong negative bias of the emulator (initialized with CS2SMOS) is itself evidence of this distribution shift. Notably, in the twin experiment, the scale of the bias error in Fig. 5b is much smaller. We added this point after L.300: ‘NeXtSIM, on which the emulator is trained, has a different SIT distribution than CS2SMOS, which could explain this phenomenon, as it can be seen with the different order of magnitude of the bias between the emulator initialized with neXtSIM fields in Fig.5b) (around 0.02 m) and the emulator initialized with CS2SMOS (around 0.1 m).’

RC: 90. Line 301: Do the percentages refer to the IIEE at the end of the forecast, or at some other lead time?

AR: All numerical comparisons are made at the end of the forecast, at the same time as neXtSIM-F. In line 276 we say: "It produces a 9-day forecast, which we will use for comparison with our data assimilation scheme." We modified this sentence to "It produces a 9-day forecast, which we will use systematically for numerical comparison with our data assimilation scheme."

RC: 91. Legend of Figure 6: The use of "ECMWF forecast" is misleading, as you are only using ECMWF fields to force your SIT emulator.

AR: Thank you for your remark, we removed the ECMWF forecast in the legend, as we explained the use of atmospheric forecast in the legend.

RC: 92. Caption of Figure 6: Please specify in the caption that the "RMSE values from neXtSIMF" are RMSEs of 9-day forecasts.

AR: Thank you for your remark. We changed the caption: "The RMSE values from neXtSIM-F, corresponding to 9 day forecast, are displayed as blue dots and should be compared with the end of each corresponding dashed line."

RC: 93. Figure 7: The graphs are a bit messy. It is sometimes difficult to distinguish between solid purple lines and dashed purple lines.

AR: Thank you, We adjusted the linestyle to improve readability, see Fig. 10 of the present document.

RC: 94. Figure 8, top-right panel: The heading "Analysis" is misleading, as you say in the caption that it is a 9-day forecast from the analysis.

AR: Thank you, we changed this heading to "4D-Var forecast"

RC: 95. Caption of Figure 8: Are the CS2SMOS observations shown in the bottom-left panel of the figure also valid on 26 March 2021, or they span over a larger date range?

AR: We use CS2SMOS observations, which are available daily with a rolling 7-day temporal average. As a result, this observation is only valid on March 26, 2021, but is strongly correlated with observations from the surrounding days.

RC: 96. Line 320: "speculate" might be more appropriate than "infer" if it is not backed by evidence.

AR: Thank you for your remark, we changed the wording. The sentence became "However, we can speculate that improving the emulator's quality—addressing both bias and smoothing issues—would result in more accurate 4D-Var analyses."

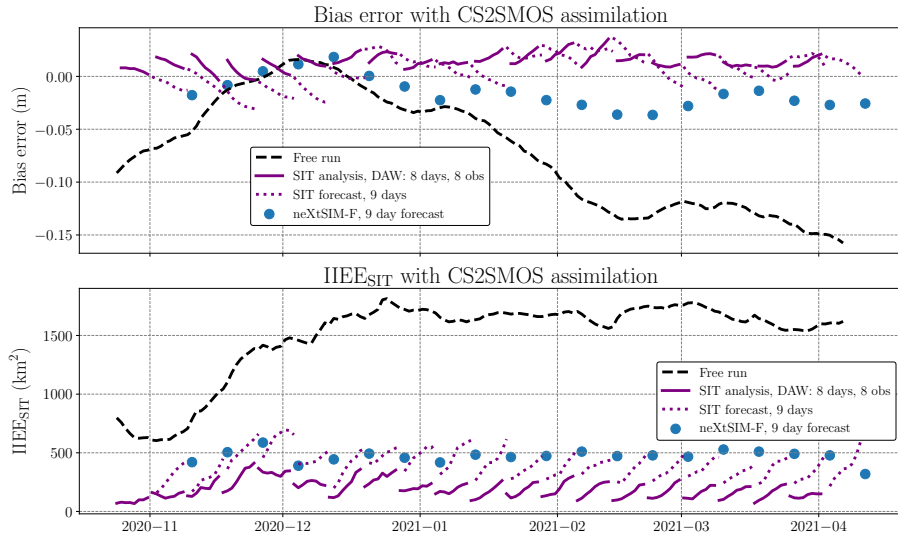


Figure 10: Bias (upper panel) and IIEE (lower panel) results for CS2SMOS assimilation across several DAWs throughout the full CS2SMOS observation period in 2020-2021 are shown. The black dashed line represents the free run of the emulator through all cycles, initialized with a SIT field from October 2018. The solid purple line corresponds to the analysis of the 4D-Var over the DAW, while the associated dotted purple lines represent the additional forecasts using ECMWF atmospheric forecasts for 9 days. The bias errors and IIEE from neXtSIM-F, corresponding to 9 day forecast, are displayed as blue dots and should be compared with the end of each corresponding dashed line. All bias errors and IIEE are computed with CS2SMOS considered as the truth.

RC: 97. Line 323: It would be good to explain what “NVIDIA A100 SXM4 80 GB GPU” means. This string of letters and numbers seems a bit cryptic.

AR: Since the following paragraph describes computation time, we decided to simply provide the GPU configuration. Computation was performed on an NVIDIA GPU from the Ampere generation (A100) with 80 GB of RAM. We kept the sentence as it is, as it may be of interest to those familiar with GPU computing.

RC: 98. Line 325: What contributes towards the time for “gradient computation”? Apart from the running the adjoint of the emulator, is there anything else?

AR: The gradient computation corresponds to computing the gradient of the cost function. This process also involves reshaping the one-dimensional vector of assimilated grid cells into the two-dimensional field required to run the emulator, as well as computing the cost function terms. Note that the emulator adjoint is computed eight times in each DAW. We added a sentence after L325: ‘The gradient computation process involves the computation of the emulator adjoint, and also reshaping the one-dimensional vector of assimilated grid cells into the two-dimensional field required to run the emulator, as well as computing the cost function terms gradients. Note that the emulator adjoint is computed eight times in each DAW.’

RC: 99. Lines 328 – 329: You say that “fewer iterations of the L-BFGS optimizer are required on average for 4D-Var-EOF to achieve the analysis at each cycle”. Do you have evidence demonstrating this?

AR: We state this fact in this paragraph based on our experiments. But you can see this from the important time difference between the 4D-Var-diag and the 4D-Var-EOF in Line 324.

RC: 100. Lines 329 – 330: You mention about the conditioning of the minimization. I would like to see how the dimensionality of the problem plays a role here, especially with small m in the EOF case (see also Minor Comment 122). Reducing the number of EOFs, as you indicate in line 335, could reduce computational time by having fewer iterations in the minimization before reaching convergence. (Theoretically, you should be able to reach the actual minimum by m iterations, when $m < N_z = 8871$.)

AR: Thank you for your remark. Please find in Fig.11 the evolution of the number of iterations of the L-BFGS minimization for all 45 cycles as a function of the truncation index. We observe the expected constant growing behavior. While we provided the analysis based on the truncation index in the Appendix. Our main focus in this work was not to find the

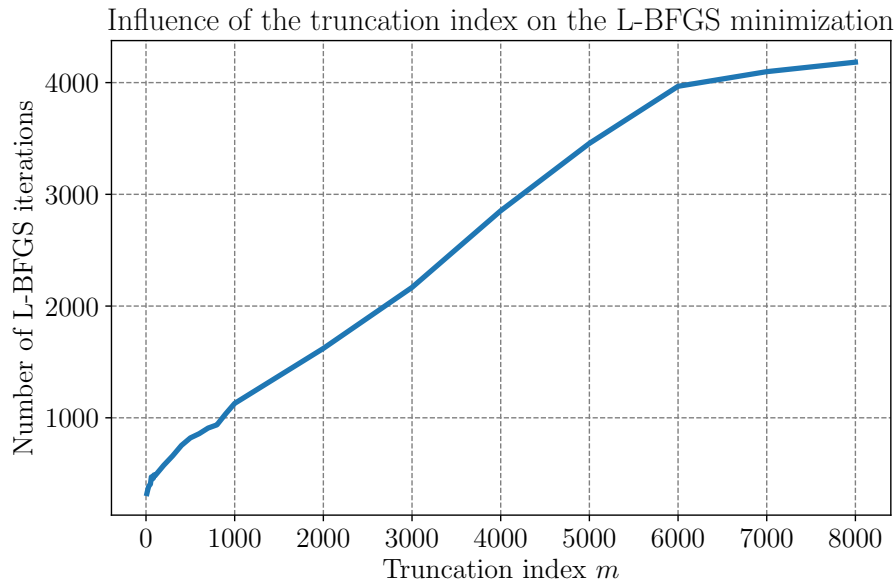


Figure 11: Evolution of the number of L-BFGS iterations as a function of the truncation index.

optimal m value but rather to simply improve the conditioning of the minimization by working in the EOFs space. As such, working with a high truncation index was a choice we made early on.

RC: 101. Lines 336 – 337: It is true that the RMSE increases when m decreases. However, according to Figure B2, $m = 3000$ and $m = 7000$ don't make a significant difference in RMSE.

AR: Thank you for your remark. We slightly updated the sentence to: "On the other side, it might lead to a loss of the small scale information, which can also be observed in Fig.B2, where the RMSE increases as the truncation index decreases, down to a certain threshold."

RC: 102. Line 339: It is not clear that the notation χ_p^2 refers to the chi-squared distribution / test with p degrees of freedom.

AR: We defined the p degrees of freedom: "Two versions are evaluated: a more often used constant model inflation and an adaptive one, based on the χ_p^2 (with p degrees of freedom) diagnostic". But we believe that the χ_p^2 literally translate to chi-squared diagnostic.

RC: 103. Line 344: "in terms of" instead of "in term of".

AR: Thank you, we corrected this.

RC: 104. Lines 345 – 346: a. I think changing the observation frequency is scientifically interesting, so the material shouldn't be relegated into an appendix. I would suggest that the additional experiments and results be discussed in more detail in the main text. **b.** It is better to use "observation frequencies" instead of just "frequencies", as the latter could mean wave frequencies (it doesn't make sense in the current context but could still create confusion for the reader). See also Minor Comment 34.

AR: a. We moved this figure to the main part of the manuscript. b. Thank you; we have corrected the wording, as mentioned in Minor Comment 34.

RC: 105. Line 346: Why do you say that current satellites can provide spatially and temporally sparse observations?

AR: With the sentence "However, current satellite data either provide spatially and temporally sparse observations, or smooth, time averaged full coverage, which introduces inherent time correlations" we were referring to typically Cryosat-2 observations. We modified the sentence to: "However, current satellite data either provide spatially and temporally sparse observations (like Cryosat-2) or smooth, time-averaged full coverage products, which introduces inherent time correlations (like CS2SMOS)."

RC: 106. Line 347: The observations themselves being correlated in time isn't a problem. It is an issue only if the observation errors are correlated.

AR: Thank you for your remark, we changed the wording of the sentence: "However, current satellite data either provide spatially and temporally sparse observations (like Cryosat-2) or smooth, time-averaged full coverage products, which introduces inherent correlations between the observations errors (like CS2SMOS)."

RC: 107. Lines 347 – 350: What is the significance of mentioning "It is important to note... outside the average observation window" and "there is currently no SIT retrieval dataset that provides daily, non-smoothed, and non-time-correlated SIT measurements"?

AR: We put this sentence to justify the fact that we use assimilated observations as our ground truth. We modified the last sentence to "Yet, in the case of real observations, there is currently no SIT retrieval dataset that provides daily, non-smoothed, and non-time-correlated SIT measurements to validate our experiments against."

RC: 108. Line 351: Localised dynamics is not the same as having less-smooth fields / more small scale features. I would interpret "localised dynamics" as having dynamical interactions that are

local. This is not easy to quantify, so I suggest avoiding this terminology.

AR: Thank you, we changed the sentence to "Let us note that neXtSIM, as well as neXtSIM-F exhibits more small-scale features than our emulator, which tends to smooth out the fields."

RC: 109. Lines 363 – 364: The gradient norm changing by more than an order of magnitude is not a problem. In the extreme case where you converge to the actual minimum, the gradient there is zero, so there is an infinite order of magnitude difference compared to the initial gradient norm.

AR: Indeed, but we meant to state this as an observed case and not a 'problem'.

RC: 110. Lines 371 – 373: Do you have a reference for the statement "In deep learning, the high number of degrees of freedom often results in poor-quality gradients"? Also, from a theoretical point of view, how does the background-error term regularise the emulator's noisy gradient?

AR: Thank you for your remark. We changed the wording of the sentence to "In deep learning, the high number of degrees of freedom often results in noisier gradients (Sitzmann et al., 2020)".

RC: 111. Lines 376 – 377 and Appendix C4: There is also another important test to check that the adjoint has been coded up properly. For M^T being the adjoint of M , $\langle \mathbf{x}, M^T \mathbf{y} \rangle$ should be equal to $\langle M \mathbf{x}, \mathbf{y} \rangle$ equal to $\langle \cdot \rangle$ (up to machine precision) for any vectors \mathbf{x} and \mathbf{y} of appropriate size, where $\langle \cdot, \cdot \rangle$ is the inner product.

AR: Thank you for your remark. Note that we haven't 'coded' the adjoint itself, as it is generated automatically with automatic differentiation. We agree that this test is pertinent; however, since we already have two conclusive tests for adjoint validation, we have decided to omit it in the paper.

RC: 112. Lines 387 – 388: For observations, a more conventional way is to thin the observations before assimilating them. Then the observations would not be regarded as a smooth field, but discrete data points.

AR: Thank you for your remark. What we meant here is that the more local dynamics are currently not available in Arctic-wide SIT observations. We removed the sentence.

RC: 113. Line 410 (Equation A1): N_s has not been defined. Please also make the equation and notations consistent with Equations 12 and 13. See also Minor Comment 48.

AR: We define N_s as the number of trajectories just above equation A1. "The root-mean-squared error (RMSE) between the prediction $\mathbf{x}_{n+k\Delta t}^f$ and the simulation $\mathbf{x}_{n+k\Delta t}^t$ is computed over all pixels (i, j) of the field of size (N_x, N_y) , for each sample n of the validation set containing N_s trajectories, initialized at time t_n ". We added its numerical value. We updated the functions for consistency with Eqs.12 and 13. Specifically removing the bold for \mathbf{x} to state $x_{i,j}$ and moving the time between the parenthesis:

$$\text{RMSE}(k) = \frac{1}{N_s} \sum_{n=1}^{N_s} \sqrt{\frac{1}{N_x \cdot N_y} \sum_{i,j}^{N_x, N_y} (\mathbf{x}_{i,j}^f(t_n + k\Delta t) - \mathbf{x}_{i,j}^t(t_n + k\Delta t))^2}, \quad (2)$$

and

$$\text{bias}(k) = \frac{1}{N_s} \sum_{n=1}^{N_s} \frac{1}{N_x \cdot N_y} \sum_{i,j}^{N_x, N_y} (\mathbf{x}_{i,j}^f(t_n + k\Delta t) - \mathbf{x}_{i,j}^t(t_n + k\Delta t)). \quad (3)$$

Referring to Minor Comment 48, it is common to present results of the emulator as a function of time.

RC: 114. Line 414: It is important to mention at a more prominent place that the outputs of the neural network model are forecast increments instead of the forecast field. Also please explain why this choice is made.

AR: Thank you for your remark. This choice is already explicitly mention in the main text with Eqs. 4a and 4b as well as the sentences Lines 141-143. Note that f_θ is an intermediate emulator. This idea is to simplify the emulator training by focusing on the differences that happens in 12 h instead of predicting directly the full state (where the small changes would be harder to predict). By splitting the full learning, we ensure that 1- the emulator learns the SIT evolution and then, 2- respects the positiveness constraint. We changed the paragraph around Lines 140 - 150: "The surrogate model g_θ predicts the full sea-ice thickness $\tilde{\mathbf{x}}_{t+\Delta t}$ with a $\Delta t = 12$ h lead time. The neural network f_θ with its weights and biases θ is trained to predict the evolution of the SIT after $\Delta t = 12$ h ($\tilde{\mathbf{x}}_{t+\Delta t} - \tilde{\mathbf{x}}_t$) based on the initial conditions $\tilde{\mathbf{x}}_t$ and given atmospheric forcings \mathbf{F} . Added to $\tilde{\mathbf{x}}_t$, this results in the prediction of the full SIT field $\tilde{\mathbf{x}}_{t+\Delta t}$,

$$\tilde{\mathbf{x}}_{t+\Delta t} = g_\theta(\tilde{\mathbf{x}}_t, \mathbf{F}_t) \quad (4a)$$

$$= \text{Relu}(\tilde{\mathbf{x}}_t + f_\theta(\tilde{\mathbf{x}}_t, \mathbf{F}_t)), \quad (4b)$$

with the point-wise activation function, $\text{Relu}(\tilde{\mathbf{x}}) = \max(\text{SIT}_{\min}, \tilde{\mathbf{x}})$, limiting the output to the lower physical bound in the normalized space. Note that f_θ is an intermediate emulator. This idea is to simplify the emulator training by focusing on the differences that happens in 12 h instead of predicting directly the full state, where the small changes would be harder

to predict as done in Durand et al. (2024). By splitting the full learning, we ensure that firstly the emulator learns the SIT evolution and secondly g_θ respects the positiveness constraint”

RC: 115. Lines 415 – 416: I am under the impression that the mean and standard deviation in Equation 1 are those for the full field, but here you discuss the mean and standard deviation of the forecast increment. Could you please clarify?

AR: Thank you for your remark. We believe the sentence to be clear enough. μ_{out} and σ_{out} corresponds to the normalization values for \mathbf{y}_t , which corresponds to the 12 hours SIT difference. As such while computed as in Eq. (1), their values differs from μ_{SIT} and σ_{SIT} .

RC: 116. Algorithms in appendices (in general): Please ensure that all notations in the presented algorithms are defined. While a specialist reader may be able to guess what they are, the presentation of the algorithms needs to be accessible to a general audience.

AR: Thank you for your remark. We checked it. For Alg 1. specifically, all notations were already introduced except σ which will be removed and change by ReLU, introduced in Sec 3.

RC: 117. Algorithm A1: It doesn’t seem clear to me how the individual lines in the algorithm presented here relate to Equations 4a and 4b, especially since there are normalised and unnormalised fields involved here. A concise, worded presentation of the algorithm would help clarify how it works.

AR: Thank you for your remark. The algorithm corresponds exactly to equation 4b) but detailing all the normalization and denormalization processes involved (because inputs and outputs of neural network are normalized). We stated this clearly in the caption ”Full-state surrogate model g_θ mapping from $\tilde{\mathbf{x}}_t$ to $\tilde{\mathbf{x}}_{t+\Delta t}$, using the previously trained f_θ . This algorithm describes how the state $\tilde{\mathbf{x}}_{t+\Delta t}$ is obtained by the application of the fine-tuned f_θ neural network as defined in Eq.(4).”

RC: 118. Caption of Figure A1: Could you explain what the “Training and validation losses” are? Also, there are two blue lines per panel, so it is not clear what “Blue line corresponds to the training of g_θ ” is.

AR: We shortly described their roles in the appendix in Lines 417-418: ”The training loss measures how well the emulator fits the training data, while the validation loss assesses its performance on unseen data to detect overfitting and ensure generalization.” We changed the caption to ”Light blue and light green lines in transparency indicate the validation losses.” to clarify the different lines.

RC: 119. Figure A2: Do you know why there remains a bias in the g_θ forecast?

AR: The emulator is trained to minimize MSE, not bias. The results presented here are from the test dataset, which the emulator has never seen. There is no guarantee that the bias would be close to 0 m. Achieving an improvement in bias through g_θ is already a positive outcome and was not guaranteed.

RC: 120. Line 426: Is the “number of state[s] in the ensemble” equivalent to the ensemble size?

AR: Yes the number of states in the ensemble corresponds to the ensemble size.

RC: 121. Line 435: It is good to describe how the EOFs are ordered.

AR: Thank you for your remark. We added 'In practice, this involves limiting the projection of the orthonormal matrix to the first m EOFs, ordered by explained variance,...'

RC: 122. Lines 442 – 443: The choice of $m = 7000$ seems quite large to me. In the limit of large m ($m \rightarrow 8871$), you would have $\mathbf{B} = \varphi\varphi^\top = \mathbf{I}$ due to the orthogonal nature of the EOFs. This means you would recover the 4D-Var-diag case, up to the σ_b^2 scaling. Usually, people look at only a small number of EOFs in order to reduce the dimensionality of the problem without compromising too much the quality of the results, so I find the choice of $m = 7000$ quite intriguing. What would you like to achieve here? Do you want to get to a close approximation to the 4D-Var-diag case? Figure B1 shows that with $m = 4$ you can already explain about 73% of the variance. How different would this be in the $m = 7000$ case? (See also Minor Comment 100.)

AR: Thank you for your remark. The idea is not to approximate the 4D-Var-diag case, but rather to do the minimization in a different subspace where we have a better representation of the cross-variances between grid-cells. As you can see in Fig. B2, if we are only selecting a small number of EOFs in the minimization, we obtain worse results than by selecting a high truncation index. The simple fact that we obtain a $\sim 15\%$ improvement in terms of RMSE, between the two types of 4D-Var indicates that we are not approximating the 4D-Var-diag case with our 4D-Var-EOF with a high truncation index. While the choice of the truncation index between $m = 5000$ and $m = 8871$ is arguable, we made a choice, based on Fig B2 and stick to it throughout the manuscript.

RC: 123. Line 446: Is there a reason that you present the algorithm for the 4D-Var-EOF case only?

AR: As mention in one of your major comment 'there is too much content in the appendix.' We decided to only display one algorithm, as both are fairly similar. The code are publicly available in case someone wants to see exactly the differences between the two algorithms.

RC: 124. Lines 448 – 449: According to Equation 11 and Algorithm C1, the H operator is operated on x space instead of w space, which is inconsistent with the statement “transformed back into the affine space of the EOFs”.

AR: Thank you for your remark, there was indeed a mistake, we corrected the sentence which was not ordered correctly ”The state w_0 is mapped back to the physical space, forecasted throughout the DAW using the emulator, where the observation term of the cost function is computed and then transformed back into the affine space of the EOFs.”

RC: 125. Algorithm C1: What are N_f ; and θ , and why is there a third argument to the function $g_\theta^{N_f}$ when there is only one in the $g_\theta^{N_f}$ in Equation 4a?

AR: Thank you for your remark. We modified the Equation 4.a to add the forcings (which are present in f_θ and by extension automatically in g_θ). θ is defined in Line 142, its the standard denomination for the parameters of the neural network (hence g_θ). N_f is defined Line 169, it corresponds to the number of application of the emulator between two observations. We recalled it there. We removed the θ in the g_θ for consistency.

RC: 126. Algorithm C2: a. The algorithm is not about the minimisation itself (which is contained within one line, the one with “L-BFGS”), but about the cycling of the 4D-Var scheme, so the heading may be inappropriate. b. How is the “loss” in the L-BFGS line related to the cost function \mathcal{J} in Algorithm C1? c. How do you ensure that the analysis (incremented from the background) stays within physical bounds? d. Would you consider combining Algorithms C1 and C2 and presenting them as one algorithm ?

AR: a. Thank you, we corrected the heading accordingly ”Wrapper for cycling the 4D-Var-EOF”. b) This is exactly \mathcal{J} , we stressed this point: ”Wrapper for cycling the 4D-Var-EOF, using the loss (\mathcal{J}) defined in Alg. C1.”, c) In the EOF case, we cannot ensure the positiveness of the analysis for the beginning of the DAW. Afterwards, since we apply the emulator in the DAW and the forecast, we ensure the physical constraint automatically. Note that in the 4D-Var-diag case, we do constrain the analysis within physical bounds, as stated in Line 452. d) We believe it is easier to read the algorithms this way. It also makes sense as it is closer to how the code is effectively implemented.

RC: 127. Lines 450 – 454: You presented two stopping criteria for the minimisation. Does it stop when either criterion is met or when both are met? Failing to meet those criteria, what is the maximum number of iterations to be run?

AR: It is stopped when one criterion is met. As mentioned in Line 455, in our case only the f_{tol} criterion is met. We never faced the case where the criteria was never met, and we actually never defined a max iteration criterion in practice. We added the following sentence after L454: "Note that we did not define a maximal number of iterations for the L-BFGS-B and the criteria f_{tol} was systematically reached."

RC: 128. Line 457: Perhaps "Cost function diagnostics" is a better heading for Appendix C2.

AR: Thank you, we corrected it.

RC: 129. Line 459, regarding the words "the seasonality... can also be observed in the cost function": This needs to be elaborated further. It is not clear how the horizontal axis in Figure C1 translates to the timing of the year.

AR: One cycle corresponds to 16 days, which allows us to interpret the timing of events. We removed the word seasonality to avoid implying a specific time of year: 'The RMSE increase observed in Fig. 5 can also be observed in the cost function evolution. The increase in May comes after 9-10 cycles, which corresponds to 2 cycles before the time where the observation cost function decreases and the background cost function increases and becomes predominant.'

RC: 130. Lines 462 – 463: What do you mean by the "adaptive background strategy", and do you mean Appendix E instead of Section E?

AR: Yes, we meant Appendix E. and we changed it. We explain the adaptive background strategy in Appendix E.

RC: 131. Figures C1 and C2: Do the figures refer to the 4D-Var-diaq or 4D-Var-EOF case, and for which type of observations?

AR: Thank you for your remark. We put more details: Caption for Figure C1 "Cost function minimization with L-BFGS optimizer across all cycles for 4D-Var-EOF with log-normal noise, the total cost function is shown in blue, this term can be decomposed with the background loss term (green term) and the observation loss term (orange curve). Note that the y-axis is in log scale." and caption for Figure C2: "Evolution of the gradient under minimization during the 4D-Var-diaq minimization. In left panel is displayed the gradient at the end of the first minimization and on the right panel is outlined the gradient at the end of the last cycle minimization. In this case, observations are built with log-normal noise."

RC: 132. Figure C1: Is there a reason to show the evolution of cost function values within a minimisation run? I think you can simply show the analysis J , J_b and J_o (the final values at the end of the minimisation) as functions of the cycle time.

AR: Thank you for your remark. We believe both pieces of information are valuable. Our figure also shows that the number of iterations varies over time. In particular, when the background cost function term increases, the number of L-BFGS iterations also increases.

RC: 133. Figure C2: It is not clear what the “first” and “last” in the caption refer to. This is also inconsistent with the text, which says “the beginning and the end of the DAW” (line 465). I don’t think much could be said about the evolution of the cost function gradient throughout the DAW, as you start with zero gradient at the end of the window and add contributions to it as you run the adjoint model backwards in time.

AR: We apologize, we meant ‘cycle’ instead of ‘DAW’. We clarified the caption: “Evolution of the gradient during the 4D-Var-diag minimization. The left panel shows the gradient at the end of the first minimization of one cycle (cycle 31), while the right panel displays the gradient at the end of the last minimization of the same cycle, when the minimization criterion is reached. In this case, observations are constructed with log-normal noise.”

RC: 134. Line 485: Do you know why the residual errors increase for $\epsilon < 10^{-7}$

AR: We do not know why it increases. But note that there are many computations taking place when calculating the losses and a growing number of numerical errors could explain this increase.

RC: 135. Line 489: Please define what M is. I think so far you have only described about the (nonlinear) emulator and the (linear) adjoint, but not a linearised version of the model that is usually denoted by M.

AR: Thank you for your remark. We changed the notation in this part. M was in the case referring to the emulator g_θ to evaluate its gradient.

RC: 136. Line 493: Is there a reason to show Equation C8c? It is not used in the test described below.

AR: Yes, it is true, we put it to show to the reader the point you were mentioning also in Minor Comment 111. But we removed this line.

RC: 137. Caption of Figures C3 and C4: a. The $\mathcal{O}(\|\epsilon\|)$ line is not an “evolution”, but just an identity line on the graph. b. In Figure C4, this line is in red, not blue, and what has been referred to

as the “Black line” in the caption of Figure C4 should have been the blue line.

AR: Thank you, we corrected both points. Fig. C3 caption: “Logarithm of the absolute value of $\mathcal{I}(\epsilon)$ for several values of ϵ . Black line corresponds to the choice of a random perturbation of the cost function, with 10 experiments performed. Red line corresponds to a perturbation inside the Central Arctic region, with 5 experiments performed. Green line corresponds to a perturbation in the MIZ, with 5 experiments performed. Blue line corresponds to the identity $\mathcal{O}(\|\epsilon\|)$.” Fig. C4 caption: “Logarithm of the absolute value of $\mathcal{I}(\epsilon)$ for several values of ϵ . Blue line corresponds to the choice of a random perturbation of the cost function, with 10 experiments performed. Orange line corresponds to a perturbation inside the Central Arctic region, with 5 experiments performed. Green line corresponds to a perturbation in the MIZ, with 5 experiments performed. Red line corresponds to the identity $\mathcal{O}(\|\epsilon\|)$.”

RC: 138. Line 495: You said earlier (line 488) that you want to test the adjoint, but Equation C9 doesn’t have M^\top in it. See also Minor Comment 111.

AR: Thank you for your remark. The wording of the section heading and line 488 might be inappropriate. We want to test the gradient of the emulator rather than the adjoint. We changed the section heading ‘Tests of the emulator and the cost function gradients’ and line 488: “To test the gradient of the emulator, we evaluate the test function”. The correctness check defined line 495 indeed evaluate the gradient of the emulator.

RC: 139. Figure D1: a. Is this figure for the twin experiment or the CS2SMOS experiment? b. In the legend, what does “LN” mean?

AR: Thank you for your remark. We updated the caption for more details: “2018 RMSE of 4D–Var–EOF depending on the number of observations in every assimilation cycle in the twin experiment case. Curves indicated the RMSE of the 4D–Var analysis with regard to neXtSIM SIT, with 2 observations per cycle (purple curve), 4 observations per cycle (blue curve), 8 observations per cycle (gray curve) and 16 observations per cycle (green curve). The dashed grey line correspond to the averaged RMSE between log-normal (LN) noise and the truth.”

RC: 140. Lines 512 – 513: I am not sure about this sentence. What is the element of (Gaussian) randomness that gives rise to the chi-squared distribution? Also, why would the inflation of the background-error term depend on the number of observations in the observation error term?

AR: We refer to Michel (2014) for the theoretical justification between the Gaussian randomness and links to the χ^2 distribution.

RC: 141. Line 514: Why is $p = 8871 * N_{obs}/2$.

AR: In Michel (2014), they explain that $2\mathcal{J} \simeq p$ with p the degree of freedom of the observations. By dividing this equation by 2 we define our own χ^2 criteria with the degree of freedom equal to the number of grid-cells (8871) times the number of observations in a DAW (N_{obs}) times divided by 2.

RC: 142. Line 517: When you speak about DOF, do you refer to the DOF of the chi-squared distribution, or the dimensionality of the cost function?

AR: We refer to the dimensionality of the cost function. We changed the sentence to "Following the χ^2 assumption, for each cycle, the minimal value of the cost function \mathcal{J} should in average be equal to the number of degree of freedom (DOF), which is equal to p " since the $/2$ might be confusing.

RC: 143. Lines 519 – 520: a. Could you please explain the motivation behind the formulation in Equation E2 and explain why the adaptive inflation factor needs to be multiplied by λ_m ? b. What are \mathcal{J}_n^b and \mathcal{J}_n^o ?

AR: a) We have the inflated cost function

$$\mathcal{J}_{\text{inflated}} = \mathcal{J}_o + \frac{1}{\lambda^2} \mathcal{J}_b. \quad (5)$$

and under the χ^2 assumption,

$$\mathcal{J}_{\text{inflated}} \sim p. \quad (6)$$

The combination of those two equations leads to Eq. E2. Then, by multiplying the adaptive inflation with the model inflation we first ease the comparison with the constant inflation scheme, and secondly still takes into account the inflation term associated with the model error. b) The \mathcal{J}_n^b and \mathcal{J}_n^o terms correspond to \mathcal{J}^b and \mathcal{J}^o for each cycle n . We stressed this point: 'where \mathcal{J}_b^n corresponds to the background term of the cost function at cycle n and \mathcal{J}_o^n corresponds to the observation term of the cost function at cycle n .'

RC: 144. Lines 524 – 525: The terminology “multiplicative addition” is ambiguous: it should either multiply or add, but not both.

AR: Thank you for your remark, we changed the sentence: 'Two schemes are evaluated, a constant inflation, only modeled by λ_m and an adaptive inflation scheme with the multiplication of λ_a based on the χ^2 estimation of \mathcal{J} and λ_m .'

RC: 145. Lines 526 – 527: According to Figure E1, when the inflation factor is larger than 1 in the constant inflation case, the mRMSE is larger, so I am not sure why you say it yields better results.

AR: Thank you for your remark, we added more details: "Using an inflation scheme ($\lambda_{\text{inf}} \neq 1$) yields better results in terms of mRMSE in both cases if we correctly tune λ_m "

RC: 146. Lines 528 – 529: For the constant inflation case, using an inflation factor above 1 means emphasising the \mathcal{J}_o term more and not the \mathcal{J}_b term, as you are assigning more error to the background. Hence, according to Figure E1, it is the emphasising of \mathcal{J}_o that undermines the performance of 4D-Var, not the emphasising of \mathcal{J}_b

AR: Thank you for seeing the typo, we indeed meant \mathcal{J}_o , we corrected it.

RC: 147. Lines 534 – 535: Do you know why you improve the results at beginning of the DAW and worsen them at the end of the DAW when adaptive inflation is used?

AR: We do not know, but we felt it was worth mentioning since it appears to be systematic.

RC: 148. Line 536: I would interpret "4D-Var is struggling" as having convergence issues in the minimisation, which does not necessarily imply the largeness of the analysis RMSE

AR: Thank you for your remark, we changed the wording of the sentence: "Let us also note that the major gain from the inflation come from March to July, which corresponds to the period where the RMSE of the analysis \mathbf{x}_a is higher"

References

- Caya, A., Buehner, M., and Carrieres, T. (2010). Analysis and forecasting of sea ice conditions with three-dimensional variational data assimilation and a coupled ice–ocean model. *Journal of Atmospheric and Oceanic Technology*, 27(2):353–369.
- Desroziers, G., Berre, L., Chapnik, B., and Poli, P. (2005). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3385–3396.
- Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W. (2012). The operational sea surface temperature and sea ice analysis (ostia) system. *Remote Sensing of Environment*, 116:140–158.
- Durand, C., Finn, T. S., Farchi, A., Bocquet, M., Boutin, G., and Ólason, E. (2024). Data-driven surrogate modeling of high-resolution sea-ice thickness in the arctic. *The Cryosphere*, 18(4):1791–1815.
- 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. (2015). 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(12):8327–8345.
- Hibler, W. D. (1979). A dynamic thermodynamic sea ice model. *Journal of Physical Oceanography*, 9(4):815–846.
- Hunke, E. C. and Dukowicz, J. K. (1997). An elastic–viscous–plastic model for sea ice dynamics. *Journal of Physical Oceanography*, 27(9):1849–1867.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.
- Ji, Q., Zhu, X., Wang, H., Liu, G., Gao, S., Ji, X., and Xu, Q. (2015). Assimilating operational sst and sea ice analysis data into an operational circulation model for the coastal seas of china. *Acta Oceanologica Sinica*, 34(7):54–64.
- Koldunov, N. V., Köhl, A., Serra, N., and Stammer, D. (2017). Sea ice assimilation into a coupled ocean–sea ice model using its adjoint. *The Cryosphere*, 11(5):2265–2281.
- Kurtz, N. and Harbeck, J. (2017). Cryosat-2 level 4 sea ice elevation, freeboard, and thickness, version 1.
- Lakshmivarahan, S. and Lewis, J. M. (2013). *Nudging Methods: A Critical Overview*, page 27–57. Springer Berlin Heidelberg.

- Lemieux, J., Beaudoin, C., Dupont, F., Roy, F., Smith, G. C., Shlyayeva, A., Buehner, M., Caya, A., Chen, J., Carrieres, T., Pogson, L., DeRepentigny, P., Plante, A., Pestieau, P., Pellerin, P., Ritchie, H., Garric, G., and Ferry, N. (2015). The regional ice prediction system (rips): verification of forecast sea ice concentration. *Quarterly Journal of the Royal Meteorological Society*, 142(695):632–643.
- Lindsay, R. W. and Zhang, J. (2006). Assimilation of ice concentration in an ice–ocean model. *Journal of Atmospheric and Oceanic Technology*, 23(5):742–749.
- Michel, Y. (2014). Diagnostics on the cost-function in variational assimilations for meteorological models. *Nonlinear Processes in Geophysics*, 21(1):187–199.
- Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L. (2019). Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10):4185–4219.
- Rampal, P., Bouillon, S., Ólason, E., and Morlighem, M. (2016). neXtSIM: A new Lagrangian sea ice model. *The Cryosphere*, 10(3):1055–1073.
- Ricker, R., Hendricks, S., Kaleschke, L., Tian-Kunze, X., King, J., and Haas, C. (2017). A weekly arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data. *The Cryosphere*, 11(4):1607–1623.
- Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473.
- Tian-Kunze, X., Kaleschke, L., Maaß, N., Mäkynen, M., Serra, N., Drusch, M., and Krumpfen, T. (2014). SMOS-derived thin sea ice thickness: algorithm baseline, product specifications and initial verification. *The Cryosphere*, 8(3):997–1018.
- Tietsche, S., Notz, D., Jungclaus, J. H., and Marotzke, J. (2013). Assimilation of sea-ice concentration in a global climate model – physical and statistical aspects. *Ocean Science*, 9(1):19–36.
- Toyoda, T., Fujii, Y., Yasuda, T., Usui, N., Ogawa, K., Kuragano, T., Tsujino, H., and Kamachi, M. (2015). Data assimilation of sea ice concentration into a global ocean–sea ice model with corrections for atmospheric forcing and ocean temperature fields. *Journal of Oceanography*, 72(2):235–262.
- Toyoda, T., Hirose, N., Urakawa, L. S., Tsujino, H., Nakano, H., Usui, N., Fujii, Y., Sakamoto, K., and Yamanaka, G. (2019). Effects of inclusion of adjoint sea ice rheology on backward sensitivity evolution examined using an adjoint ocean–sea ice model. *Monthly Weather Review*, 147(6):2145–2162.

- Usui, N., Wakamatsu, T., Tanaka, Y., Hirose, N., Toyoda, T., Nishikawa, S., Fujii, Y., Takatsuki, Y., Igarashi, H., Nishikawa, H., Ishikawa, Y., Kuragano, T., and Kamachi, M. (2016). Four-dimensional variational ocean reanalysis: a 30-year high-resolution dataset in the western north pacific (FORA-WNP30). *Journal of Oceanography*, 73(2):205–233.
- Wang, K., Debernard, J., Sperrevik, A. K., Isachsen, P. E., and Lavergne, T. (2013). A combined optimal interpolation and nudging scheme to assimilate osisaf sea-ice concentration into roms. *Annals of Glaciology*, 54(62):8–12.
- Williams, T., Korosov, A., Rampal, P., and Ólason, E. (2021). Presentation and evaluation of the arctic sea ice forecasting system neXtSIM-f. *The Cryosphere*, 15(7):3207–3227.
- Xie, J., Counillon, F., and Bertino, L. (2018). Impact of assimilating a merged sea-ice thickness from cryosat-2 and smos in the arctic reanalysis. *The Cryosphere*, 12(11):3671–3691.
- Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M. (2019). The ecmwf operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean Science*, 15(3):779–808.
- Ólason, E., Boutin, G., Korosov, A., Rampal, P., Williams, T., Kimmritz, M., Dansereau, V., and Samaké, A. (2022). A new brittle rheology and numerical framework for large-scale sea-ice models. *Journal of Advances in Modeling Earth Systems*, 14(8).