Response to Referee 2

for "Deep learning of subgrid-scale parametrisations for short-term forecasting of sea-ice dynamics with a Maxwell-Elasto-Brittle rheology"

Finn, T.S., Durand, C., Farchi, A., Bocquet, M., Chen, Y., Carrassi, A., Dansereau, V.

19th April 2023

RC: Reviewer Comment, AR: Author Response

RC: In this manuscript, the authors present a novel machine-learning method to correct unresolved sea-ice dynamics in simulations with low resolution. From the comparison of high and low-resolution simulations in an idealized domain neural networks are trained to predict the residual between both simulations at a certain lead time, which demonstrate promising performance. This approach aligns with recent developments in climate research, where machine learning is used to parameterize unresolved processes in low-resolution simulations. The study presents several innovative approaches to sea-ice science and provides a thorough evaluation of the performance of the trained ML algorithms. To the best of my knowledge, this paper is one of the most advanced works employing machine learning in the field of sea-ice dynamics. The presented analysis is sound and requires only a few modifications that I list below. The authors need, however, to improve the paper's presentation of the paper as it can be difficult to follow in major parts. The manuscript is overly packed with information and details that can be challenging to grasp, even with a background in sea-ice dynamics and machine learning. I strongly recommend the manuscript for publication in The Cryosphere after the authors have addressed the issues mentioned and detailed below.

AR: We thank you for the constructive feedback on our manuscript, especially on the presentation of the methods and the many detailed specific comments. In the following we will discuss the raised concerns and what we plan to change in our revised manuscript.

RC: Target audience: I think the authors should keep two audiences in mind that will be interested in this work: sea-ice scientists and ML experts. The manuscript in its current form describes the ML part, network design, and thorough evaluation of the performance of the NN in great detail. I appreciate this for reproducible science, but am also afraid that the amount of detail makes the manuscript hard to follow for readers with a sea-ice background and limited knowledge of ML. This could be addressed by shortly introducing the many ML concepts before discussing them in length and/or reducing or reorganizing the information content of the paper (which I will explain in the next point). I highly recommend reading and editing the paper through the lenses of both audiences.

AR: Exactly as you have proposed, we have both audiences in mind. Based on your review and comments of referee 1, we have seen that we might have missed the fine line between both communities by giving too much technical details. To improve the presentation for both communities, we will move the technical parts into an Appendix and explain the sea-ice model and the neural network more concisely and by rather discussing the reasoning behind different choices and components. We hope with these changes, we can satisfy readers from both communities.

RC: Readability: I had a hard time following the first half of the paper on my first read. After reading the entire manuscript and knowing the subject, I could follow it better on a second read. Therefore, I would suggest editing and maybe restructuring this part of the manuscript thoroughly. In general, the manuscript holds a lot of information in part to describe the set-up, analysis, and results in detail, but also information that is only linked but not strictly necessary for the understanding or interpretation of the paper. Especially the latter makes it hard to stay focused on the storyline. I recommend going through the paper and reconsidering which information is necessarily required. This would also give the authors more space to explain important concepts in more detail. Section 3.1 helped me a lot to understand what you are after and I definitely recommend moving it further up in the manuscript, maybe even into the introduction. I would also consider moving the description of the data generation (Section 4) before the description of the ML, which would help to understand the network design etc. Section 2 is rather long and I would consider shortening it and eventually merging it with Section 4 as both discuss the sea ice model and the simulations. Up to Section 5, I had a hard time finding a storyline to follow. Please try to emphasize your storyline there stronger and try to guide readers better.

AR: To improve the readability of the manuscript, and as you have proposed, we have decided to change the structure in presenting the methods. We will replace Section

2 by a new Section, where we explain first the problem that we try to solve in mathematical terms (former Section 3.1), second rather shortly the regional sea-ice model and the used forcing, and third the twin experiments. In Section 3, instead of a technical description, we will explain rather superficially the neural network and its reasoning. As we have moved the explanation of the forcing and of the twin experiments into Section 2, Section 4 will be more disentangled and more restricted to specific parameters etc. used to generate the data and to train the neural networks. Additionally, we will be more consistent, in the use of 'subgrid-scale parameterisation' and 'model error correction'. We believe that these changes will increase the readability of the manuscript, especially for non-specialists. Thanks for this suggestion.

RC: Lead time for update: The authors use a lead time of 10 min 8 s to update the coarse resolution model. While all other design choices have been explained in detail, this is not the case for the lead time. Why did you choose this lead time? Wouldn't you expect a shorter lead time to improve the results? With the existing twin simulations, it is straightforward to extract the residual between the truth and forecast model also at other lead times. Therefore, I strongly suggest studying also the effect of different lead times here. I would be especially interested to see if shorter lead times improve the seesaw patterns of the trajectories of the hybrid model in Figure 7.

AR: Originally, we have selected the update time quite early in the research based on considerations about the signal-to-noise ratio. In the best case, the neural network corrects model error before they have a too large influence on the forecast, which would be after each integration step. However, the neural network is not trained to take interactions between correction and sea-ice model into account. Furthermore, the predictable error is corrected by the neural network, such that the unpredictable error accumulates over time. As the network is only trained on the first correction step, this leads to a distribution shift (cf. Tab. 8 of the manuscript).
This distribution shift becomes important the more update steps we take. To underline these pros and cons, we have now made tests where we vary the update time (16 s, 80 s, 20 min), as can be seen in Figure 1 of this response (the results for 16 s are not shown, as they are even worse than for 80 s). The seesaw pattern gets hidden behind more update steps. Nevertheless, the distribution shift outweighs that model errors are corrected earlier, which makes the forecasts less performant for smaller update times. This might be a specific concern in our model setup, but is likely to be an issue related to hybrid modelling in general. As these results underline the importance of the distribution shift, we will add a Subsection in the additional results part of the Appendix where we show and shortly discuss Figure 1.

RC: Generalization: The neural networks presented in the paper are trained on a specific (idealized) model configuration, which is also a good choice

for this proof of concept. There is, however, only limited discussion of what steps are needed to use the same approach in other model configurations, especially realistic ones: do users need to train different NNs for each new model configuration, which will get very expensive as high-resolution truth simulations are required? Or can the trained weights of the kernel be applied also to different grid geometries in different configurations or could be used as starting weights to reduce the amount of training data? A discussion of these considerations would be helpful to get an impression of how feasible and flexible this approach can be applied in other model set-ups.

AR: This study has a very limited scope of giving a proof-of-concept. For us, it is evident that the way towards operational settings is rather long, given the issues of missing interactions and stochasticity. Additionally, if neural networks are trained with twin experiments, they "only" learn to imitate the model that was used to generate the truth. We rather see the potential for model error corrections coming from the inclusion of observations into the learning process, which complicates model error correction even more. To give some hope, especially related to the question of the reusability, we might be lucky in sea-ice modelling: as sea-ice exhibits multifractality/self-similarity, the same model error correction might be usable across spatial scales, or at least a good starting point for retraining on a different resolution or domain.

In the revised version of the manuscript, we will provide some more discussions upon the next steps and what is missing towards an operational setting.

RC: L1: "of"
Remove "of" as "subgrid-scale" is an adjective.

AR: We will remove "of" from the title.

RC: Abstract: I would consider rearranging the abstract, maybe shortening sentences. Might be a matter of taste, but I had a hard time following it reading it the first time.

AR: We will simplify the abstract and its language wherever we can.

RC: L5: includes important inductive biases needed for sea-ice dynamics.
Unclear what is meant by these biases.

AR: We will remove this subordinate clause from the abstract.

RC: L7: we cast the subgrid-scale parametrisation as model error correction.
Unclear, please rephrase.

AR: We will reformulate to "Instead of parameterising single processes, our goal is to correct all model variables at the same time."

RC: L11: cycling
What do you mean by cycling?

AR: We will be more specific by stating: "Correcting the forecasts every 10 minutes, the neural network can be run together with the sea-ice model, which improves the short-term forecast up to one hour."

RC: L11: physically-explainable input-to-output relation
It is not clear what is meant by "physically-explainable", please clarify.

AR: We will be more specific in what we mean by physically explainable.

RC: L16: dynamics of sea ice at an unprecedented resolution and accuracy. Please clarify what unprecedented means with respect to the resolution. All three papers use simulations with a resolution of 10km or lower, while much higher resolution sea-ice simulations have been presented. Do you mean unprecedented accuracy at the given resolution?

AR: We will rephrase to: "at an unprecedented accuracy for Arctic-wide simulations in the mesoscale with horizontal resolutions of around 10 km."

RC: L16: Elasto-Brittle. Why capitalized? Here and elsewhere

AR: We will write elasto-brittle in lowercase letters.

RC: L17: represent
Reproduce?

AR: We will change to "reproduce" as this is indeed a better wording. Thanks for the suggestion.

RC: L19: single grid cell at the mesoscale. What is meant by mesoscale here? Please clarify

AR: In the revised manuscript, we will define mesoscale in the first sentence of the introduction.

RC: L31: the mesoscale. See comment above
please define the length scale mesoscale refers to here.

AR: Same as before, mesoscale is now defined in the introduction.

RC: Figure 1. Please clarify that (a) shows the high-resolution initial conditions, but (b) and (c) the low-resolution forecasts one hour later. Why not show for all the damage after 1h forecast, so that the reader actually gets an impression if the hybrid model in (c) is closer to the high-resolution "truth" or not?

AR: The Figure shows the field for all simulations after a lead time of one hour, even in the high-resolution case. As this was not clear in the caption, we will clarify it, as shown in Figure 2 of this response. Additionally, given the feedback from Referee 1, this Figure will be changed to a sample more representative for typical situations in the dataset.

RC: L37: possibly projected
What is meant by this? Please clarify

AR: We wanted to state that we cannot use the same initial conditions across different resolutions, so we must project them. We will rephrase this sentence to clarify it.

RC: L39-40: Here, the low-resolution simulation 40 (b) misses the rapidly developed opening of sea ice in the high-resolution simulation (a).
Does this refer to the upper or the lower opening in the figure? Please clarify in the text.

AR: As the shown snapshot will change, see Figure 2 of this response, also the text will be different, and we will be more specific.

RC: L54-59: paragraph about marginal ice zone:
Does your regional model include the MIZ? To me, it looks more like pack ice with cracks. Also along leads there are sharp transitions that the NN needs to handle, so I think it is justified to present this issue here. However, please frame it in a way that fits your problem at hand.

AR: You are right that our problem is more about cracks/leads with such sharp transitions. As the model is unable to represent ice-free cases, it cannot simulate "real" marginal ice zones. We will change the description here and elsewhere in the text, being more consistent with the actual problem represented by our model.

RC: L56: jump
Step function?

AR: Step function is indeed the more accurate wording, thanks.

RC: L98: as well
Remove?

AR: "As well" will be removed.

RC: L123-124: As the nodes are shared in the first-order elements, there are more grid points for all variables that are defined as zeroth-order elements than for the velocity and forcing components.
What is the relevance of this? Could you elaborate if this is an important point needed to be considered to interpret the presented results?

AR: This reference to more grid points for variables defined by a zeroth-order discontinuous Galerkin discretisation is relevant to explain the need of a common space for the neural network. As this Section will be rewritten in the revised manuscript, and this more technical description will be moved into an Appendix, we will remove the sentence.

RC: Section 3: A deep learning based subgrid-scale parametrisation
The presentation of the machine learning tools is done very thoroughly, which I appreciate and see as valuable for reproductivity. Given that ML applications in this field of science are just emerging and there are many geophysical and climate scientists interested in advancing in this field, I am afraid that the description is presented too high level for an audience with limited knowledge of ML. To also target this part of the scientific community and broaden the audience for this paper, I recommend summarising the main parts and ideas behind it more comprehensively for readers with limited ML background at the beginning of this section. While I see this recommendation as optional, as all necessary information is given in the current draft, I want to emphasize the large beneficial value I see in adding a summary like this.

AR: As previously written, we will rewrite this Section and be more concise with the technical description in an Appendix. Thank you very much for this constructive comment.

RC: L147-149: There, linear functions combine pixel-wise (i.e., processing each element defining grid point independently) the extracted features. Each linear function is shared across all grid points for each predicted residual variable.
The linear transformation from features to residuals is not clear to me. Does this involve combining different features for each grid point, where the weights of these combinations are learned in the training? Or is it a fixed combination? Please clarify the text accordingly.

AR: We will be more specific what we mean by linear functions, as they are learnable.

RC: Figure 3.
Does the red, blue, and grey color code for arrows, boxes, and labels have a specific meaning (trainable vs fixed or similar)? If so please give some explanation.

AR: We will revise the figure, to mark learnable parts by the same colour, and we will specify our colouring code.

RC: L154: 3.1 Problem formulation
This section helps to understand our approach's goal, and I strongly suggest moving it further up (maybe even the introduction) to give the

reader a better understanding of what you try to achieve before going into the details of the model or the pipeline.

AR: This will be the first part of the new Section in the beginning of the methods.

RC: L180-182: Note, for coarse Cartesian spaces, the mapping from Cartesian space to triangular space can be non-surjective, meaning that not all triangular elements are covered by at least one Cartesian element: the pseudo-inverse is in this case rank deficient.
This is unclear to me: why should the bigger triangles of the coarse resolution simulations not be covered by the higher resolution Cartesian elements? If at all, I imagine that should be an issue of the high-resolution grid with smaller triangles. Please clarify.

AR: As the Cartesian elements are rectangular, there are cases where not every triangular element has an associated Cartesian element in the forward interpolation. Consequently, taking the pseudo-inverse of the interpolation operator for the projection, there are some triangular points that get no information, even if the triangular resolution is around 8 km and the Cartesian resolution around 4 km. This is specific to our choice of taking the pseudo-inverse instead of defining a new projection operator. This technical detail will be moved to an Appendix.

RC: L196: complete U-net architecture
Do I understand the architecture correctly that you downscale only once in your U-Net? If that is the case, the illustration in Figure 3 is misleading, as 4 down scaling steps are shown. Please clarify this and adapt the figure potentially.

AR: Yes, indeed in the shown architecture there is only one downscaling operation, we will adapt the figure.

RC: Section 4: Experimental setup
This section describes how data to train the NN is created. Consider renaming this section to e.g. "training data generation" or similar. I also would consider moving this section before the details on the ML algorithms as I feel it helps to know the data before getting introduced to the detailed methods.

AR: We will rename the section to "Data generation". By moving the twin experiment explanation part before the neural network, the ordering should be clearer, and the section more disentangled.

RC: L305: their expectation
their expected value?

AR: We will change to expected value, as this seems to be a more common name.

RC: Table 3 - Caption: MAE
Is the MAE computed at high or low resolution?
A score of one would correspond to the performance of the geophysical model forecast in the training dataset.
Do you mean the coarse resolution geophysical model forecasting the high resolution run? Please clarify.

AR: We will be more specific about low- and high-resolution in the caption.

RC: Table 3 - Caption: the afterwards used architecture and the best scores
If this refers to the hybrid model or online use of the correction, please write this. Also, consider writing "that shows" instead "and".

AR: We will specify that the architecture is used for all experiments afterwards. The "and" is chosen as the bold font is chosen independently for each score. It is by chance that all bold numbers are in-line with the hybrid model.

RC: L320: persistence forecast performs
Could you for clarity once define what you use as a persistence forecast? It might be obvious to you, but for readers outside the field, it will be helpful.

AR: We will be more specific about the persistence forecast and will define it in the beginning of the results section.

RC: L335: Such localised features
Consider adding the length scale in km if you think this finding is generalizable or holds valuable information for other processes related to sea-ice deformation.

AR: In our case, we know that the missing processes lie between the 8 km and the 4 km, but if learned from observations, it might be indeed interesting to see the length-scale of the learned model error correction. For the specific case, shown in Fig. 5, we will add one or two sentences about the features in the different resolutions.

RC: L340: a generally smoother background pattern
What is meant by this? Please clarify and rephrase.

AR: We will clarify what we mean by a generally smoother background pattern.

RC: Figure 6
I) What colormap uses a)?
II) I am wondering if also the concentration maps in the initial and forecast step would be helpful here to interpret the gradients?

AR: The colormap of (a) is somewhat arbitrary as the normalized field like used for the input into the NN is shown. Instead, we will change to the unnormalized area and show a colorbar.
To make the Figure as simple as possible, we have restricted it to only the information needed to make the point clear. We will see if we add more fields to make the gradient more explainable.

RC: L378: either
Isn't it to both instead of either or?

AR: Will be changed to "both" and "and".

RC: L380: Table ??
correct reference

AR: The reference will be corrected.

RC: L382-383: Additionally, the sensitivity is directional dependent, Fig. 6g, and exhibits localised features, Fig. 6c and i
Could you discuss these results also in the light of physical understanding that we can gain from the gradients? From both the gradients along initial and difference, we can learn about the shortcomings of the coarse resolution simulations that the NN tries to compensate for.

AR: Thank you for this nice suggestion. Our goal was to prove the point of localised features and representation of anisotropy, neglecting its physical meaning. We will add some discussion about the physical meaning along the suggested lines.

RC: L391: -1×10^{-3} **and** 1×10^{-3}
Units?

AR: This sentence is a remnant of an older version of the manuscript. As there is no longer a restriction to the correction of the sea-ice thickness, the sentence will be removed. Thanks for spotting this typo.

RC: L392: Related to optimal control theory in dynamical systems,
This is not very helpful for readers with limited background knowledge. Please elaborate more or rephrase.

AR: We will change the sentence to "As more commonly used to evaluate forecast performances, we will ..."

RC: L405-408: Additionally, for the velocity, stress, and damage, the drift towards ... the "Initial + Forecast" experiment in these variables and averaged over all nine model variables.
Unclear, please rephrase.

AR: We will rephrase the sentence to make clearer what we meant.

RC: L412: As the initial condition error increases with each update, the network corrects less and less forecast errors.
Could this effect be dampened by updating at higher frequencies?

AR: As discussed for your comment # 3, there is a trade-off. It seems that the cons outweigh the pros, and more updates even hurt the performance. As we will add the Figure to the additional results Section in the Appendix, we will refer to these results in L412.

RC: L413-419:"To show the effect of this error distribution shift, ... An averaged correlation of 1 would indicate a perfect pattern correlation."
This paragraph is hardly understandable with no background knowledge of the method "centred spatial pattern correlation". I suggest describing the principle of the method and its interpretation at the beginning of the paragraph in 1-2 sentences, before describing its specifics.

AR: We will add a short explanation of the spatial pattern correlation.

RC: L422: especially for the divergent stress
From the previous paragraph, it sounds as if a value close to 1 is favorable, but this statement reads as if a high value close to 1 for divergent stress shows a weakness of the NN. Please clarify.

AR: Considering the results for the sea-ice area, we will clarify that this shows a deficiency of the neural network but of the forecast model.

RC: L434-435: However, the parametrisation misses the development of new strains and positions the main strain at the wrong place.
This suggests that the corrections of the NN violate the brittle model physics, as highly damaged areas are usually linked to high deformation rates. Is this correct? If so, please comment on this also in the text, and if there is a way to design a network that computes corrections in accordance with the physical laws of the model.

AR: In the hybrid model, the field is damaged beyond a given deformation threshold, shown for the damaging process in the south, as in the high-resolution simulation. Only, the high deformation rates in already damaged areas are striking in the case of the hybrid model. As such high rates are unobserved for the high resolution simulation, it remains unknown if this is an unphysical behaviour or not. Because the last correction of the neural network is already more than 9 minutes ago, we would speculate and say that the model would have had time to adapt to the new situation such that the rates are physical explainable.

RC: L456-457: Therefore, using such a mapping into Cartesian space, we can apply CNNs, which can efficiently scale to larger, Arctic-wide, models.
Are you talking about Arctic-wide models on unstructured grids? Or why is the mapping needed? Please clarify the text.

AR: We will clarify that we mean Arctic-wide simulations on unstructured grids, e.g., neXtSIM.

RC: L462-463: As processes have no discretized resolution in realworld, we would have difficulties to find the right resolution for the projection in such cases.
Isn't that only an issue if you would aim to train a correction with observations? If it is a model, then you would always know the resolution of resolved scales. Please clarify

AR: Indeed, this is only an issue if we train with observations. Since in the ideal world, we learn from observations, we raise this issue here. We will disentangle the sentences and make our points clearer, also with respect to the higher-resolved truth.

RC: L462: truth
Please clarify what is meant by truth: the high-resolution simulation or something different

AR: See previous comment.

RC: L464: this argument
What argument?

AR: The argument that the optimal resolution is linked to the resolution of the processes in the "truth", either in high-resolution simulations or observations. We will clarify the sentence.

RC: L503: The only way is therefore to improve the forecast model, thereby changing its attractor.
What about updating the forecasting model at higher frequencies? Please comment.

AR: In the most extreme case, we would correct after each integration time step, which could be seen as integrated form of a subgrid-scale parameterisation. In this case, we would change the attractor, as each forecast is now a corrected forecast. The problem is rather to define the attractor of the hybrid model; one could define only the corrected states as attractor. We will make such points clearer here.

RC: L530-532: Mapping the input data into a Cartesian space that has a higher resolution than the original space, such scalable convolutional neural networks can be applied for feature extraction in sea-ice models defined on a triangular or unstructured grid.
Something is wrong with this sentence, please correct it.

AR: This sentence does make indeed no sense, we will rephrase it.

RC: L543: total deformation
On which Figure or result is the statement that the total deformation is improved? From Fig. 8 ,I would agree that the damage in the hybrid model looks closer to the high-resolution run than the uncorrected low-resolution simulation, but for total deformation, it is the other way around in my eyes.

AR: We agree that the point-wise error for the hybrid forecast in the total deformation is higher. Hence, we will rephrase it into: "... to a better representation of the damaging processes than in the forecast model without parametrisation".

RC: L552: Appendix
Table?

AR: In the current online version of the manuscript, the Table is correctly referenced?

RC: L562-564: Caused by their limited capacity, the NNs have to focus on some variables, creating an imbalance between variables, which harms the performance for other variables, like the stress in the case of "Conv ($\times 5$)".
Have you tried to train individual networks for each variable, which could balance this effect?

AR: In the initial phase of the research, we have seen that one big neural network with shared parameters performs better than nine small neural networks. Consequently, we have concentrated the research on a single big neural network. As one big network shares the features for all variables, it has to learn more general features than networks for single variables. The shared features act as regularisation which can help to reduce the overfitting. Hence, learning one network for all variables can enable the use of larger neural networks. We currently run a test for the "Conv ($\times 1$)" architecture with multiple networks and add the results to Table B1. As preliminary results, we see that learning single "Conv ($\times 1$)" networks per variable has better scores than the shared "Conv ($\times 1$)" architecture. However, the single networks have already a larger error compared to the shared "Conv ($\times 5$)" architecture, although they have in total more parameters, and they are slower to train. Additionally, the single networks have the same imbalances as evident for the shared network. Consequently, this confirms our initial results that one single big neural network for all variables performs better than one small network per variable.

RC: L565: fast NN
Is the speed of the trained U-NeXt NN an issue compared to the computational costs of the geophysical model? Or does this refers to training speeds?

AR: Compared to the geophysical model, the neural network is fast, if correctly implemented into the model, e.g., using the C++ libraries of PyTorch or TensorFlow.

However, such kind of model error correction sits on top of a geophysical model, its runtime is additive to the runtime of the geophysical model. Consequently, the faster and lightweight the model, the better. So, it is in some sense a trade-off between additional runtime and additional gain, and, for some purposes, the Conv ($\times 1$) architecture might be enough.

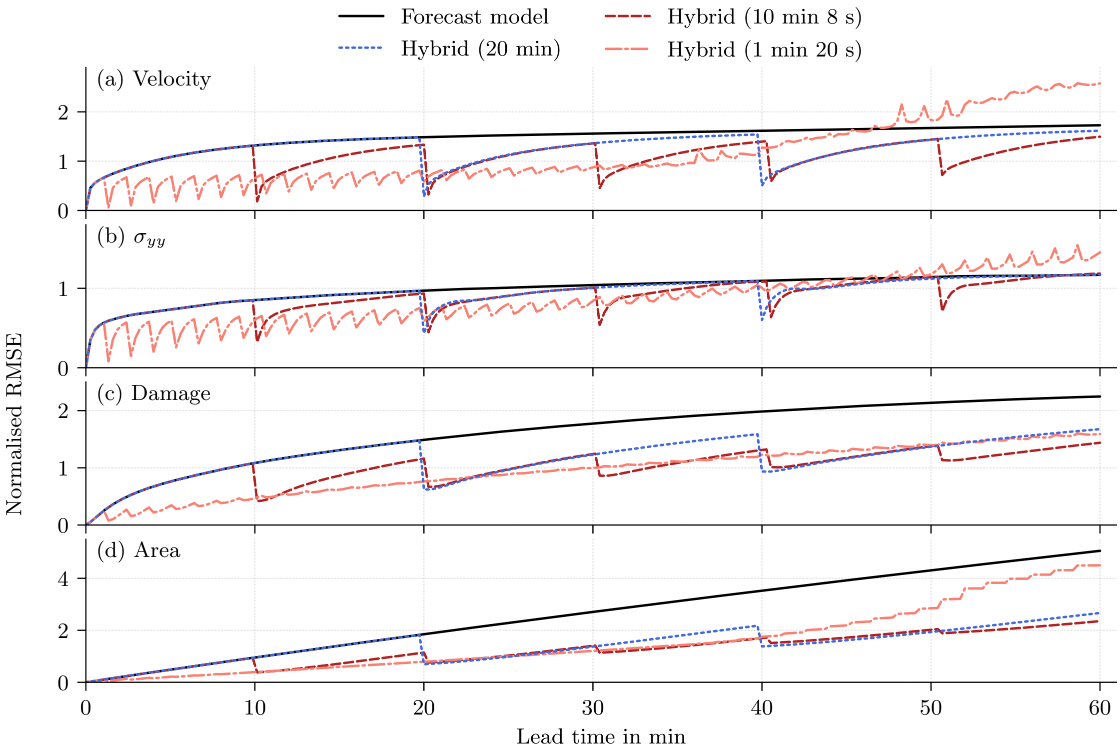

Figure 1: Normalised RMSE for (a) the velocity in y-direction, (b) the divergent stress in y-direction, (c) the damage, and (d) the sea-ice area as function of lead time on the test dataset, normalised by the expected RMSE on the training dataset for a lead time of 10 min and 8 s. In the hybrid models, the forecast is corrected after each specified lead time in brackets, and the performance is averaged over all ten random seeds

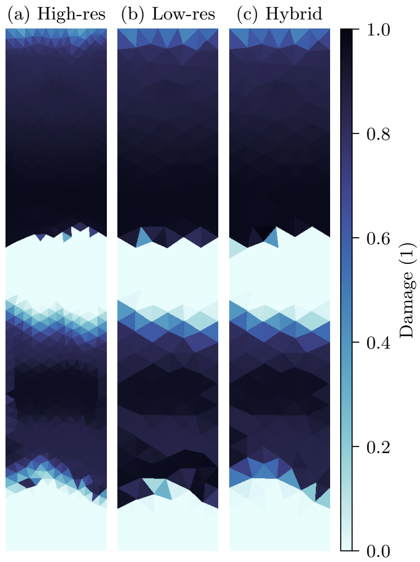

Figure 2: Snapshot of sea-ice damage after a one-hour forecast with the here-used sea-ice dynamics only-model. Shown are the high-resolution truth (a, 4 km resolution) and low-resolution forecasts (b, c). To initialise the low-resolution forecasts, the initial conditions of the high-resolution are projected into a low-resolution space with 8 km resolution. Started from these projected initial conditions, the low-resolution forecast (b) generates too much damage compared to the high-resolution field. Running the low-resolution model together with the learned model error correction (c) leads to a better representation of the damaging process, which improves the forecast by 62 % in this example.