Reviewer comment: Line 83-86. As discussed above, a neural network cannot a priori "solve the 'inverse problem' ", as is misleadingly stated here.
Response:
The wording in the original sentence 'Following the physically consistent SD, scientific machine learning (SciML) models can be used to solve the 'inverse problem',...' has been rephrased:
'Following the physically consistent SD, scientific machine learning (SciML) models can be used to approximate solutions to the 'inverse problem'...,'

Reviewer comment: Line 135. Can you characterise the uncertainties used for pruning the channels to separate error coming from the network performance and the error originating from unsuitability of the channels? If yes, how? If not, why is this approach still valid?
Response: In our experiments, using AANN models with a network structure of 13 × 7 × 13 or 12 × 6 × 12 (heuristically, the number of neurons in the middle layer being half of the other two layers), we were able to correctly identify channels requiring additional calibration. However, this approach is quite heuristic and it is difficult to isolate errors from the two sources.
When unsuitable channels are omitted from our model, the overall band-averaged deviation is significantly reduced, and the reduction can be as great as one order of magnitude: the 'restoration error' for an AANN to replicate the spectral radiance (input) can be reduced to less than 1 percent.
The validity of the approach is provided in the given reference [1]:
After being trained by the same training dataset, the aaNN works as a duplicator. If the input data are within the range of the training dataset and the shape of the spectral L_{rc} is very close to that of the training dataset, then the aaNN output will duplicate the input spectral L_{rc} with a very high precision. But if some of the input data are out of the range or the shape of the spectral L_{rc} is not included in the training dataset, then the output from the aaNN deviates significantly from the input spectral L_{rc}, i.e. the band-averaged percentage difference is larger than 5%. Therefore, by comparing the output from aaNN with the input spectral L_{rc}, we can identify pixels that are out of scope of the training dataset. This capability of the aaNN is due to the bottleneck layer in the neural network structure. The aaNN also has 3 hidden layers. The number of neurons in the first and third layers is set to equal the number of inputs. The second layer is the bottleneck layer with a much

[1] OC-SMART: A machine learning based data analysis platform for satellite ocean color sensors

smaller number of neurons, and designed such that a well trained aaNN is able to duplicate only input data available in the training dataset.

As a side note, in addition to selecting channels, the authors of OC-SMART include the aaNN as a pre-processing step in the ocean-color retrieval workflow: when satellite data are entered, the aaNN checks for pixels outside the scope of the training dataset.

Reviewer comment: Line 298. As discussed above, give information about all the networks performances.
Response: We greatly value your remark on local minima and this recommendation. The results of three other neural network (NN) models that we trained have been added to Table B1. In the meantime, Figure B1 was added to compare the histograms of the four NN and four machine learning (ML) models. Note that the validation loss (on the 20% of data excluded from model training) of all ML models was at least one magnitude greater than that of the four NN model. Figure B1 (c) and (d) illustrates how this 'under-performance' impacts final retrievals.
From Figure B1 (a) and (b), it can be seen that the four neural network models are highly comparable, at least in terms of the sampled data used to generate the histograms. In addition to Fig.B1-a, as proposed, we gathered additional independent data to illustrate the overall value ranges of the four NNs. The two panels suggest, citing your remark, that "any ambiguous solutions still give similar results or maybe there even exists an analytic inversion'. However, we recognize that this small sample is by no means exhaustive, so there may be adversarial instances that we missed. Therefore, the concluding remarks on page 34 were also revised to reflect this limitation.

Reviewer comment: Figure 4 and 6. Does this validation contain the data used for selecting the best models? (It should not!). Generally also the comparison with other products should not be made over areas that have been included in the training or selection process of the neural network models, as this would wrongly skew the results in favor of the machine learning approach.
Response: Approximately fifty percent of the data presented in Fig. 4(d) was used to determine the final model mentioned in subsequent sections.
Although this approach is not ideal, it cannot be avoided: (a) It is not possible to include the retrieval results of all NN models and compare them with the results of other approaches (melt-pond detection, MCD43, and direct-estimation method), and (b) the NN models exhibited identical loss on the

independent testing dataset (20% of the dataset generated by the radiative transfer model).
Therefore, we had to rely on *in-situ* measured data to select a 'winning model', as no other data sources were available for this purpose.

Reviewer comment: Line 591. The discussion needs to include the fact that the performance of the forward model is a hard limit on the performance of the retrieval, as sea ice in particular is known to be difficult to model accurately.
Response: We appreciate the recommendation. The relevant information is added in the Conclusion section.