Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-128-RC4, 2020

Geoscientific
Model Development
Discussions

Interactive comment on “DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations” *by* Alexander Barth et al.

Anonymous Referee #2

Received and published: 8 February 2020

Review of DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations

This study presents a novel approach of reconstructing sea surface temperatures from cloudy satellite data by making good use of modern deep learning techniques. While I believe that the study has been carefully designed and executed, I have severe problems with the paper in terms of its presentation and accuracy in writing. This study could become a high-impact publication if it were better structured and methods and outcomes were described clearer. I advise the authors to apply major revisions and seek the help of a native speaker to avoid erroneous or ambiguous statements. Below,

please find my detailed comments.

Abstract: the sentence "However, it is unclear how to handle missing data (or data with variable accuracy) in a neural network when using incomplete satellite data in the training phase." is not very clear. Perhaps rephrase as "Contrary to standard image reconstruction with neural networks, this application requires a method to handle missing data (or data with variable accuracy)."

L7: suggest to remove "essentially"

L9: what is "relatively long"? Provide a number, please.

L11: "a method to reconstruct missing data": suggest to rephrase "a previously published method", "the current standard method", "the state-of-the-art DINEOF method", or similar.

L16: what is meant by "the ocean current signal"? A signal always refers to a measurement or sensing process. Here you want to refer to a physical process in the ocean.

L17: replace "like" with "e.g."

L20: replace "sensor" with "measurement". This sentence refers to the measurement principle, not the technical instrument, which performs the measurement.

L22: "but often small scale information is filtered out because of the transient and stochastic nature of these structures." The transient and stochastic nature produces variability and is clearly not the reason why small scale information is filtered out. This is rather a result of the averaging procedures which are applied in practically all known techniques to interpolate. As this is a critical sentence for setting the stage of this study, please rethink the phrasing and provide a more precise description of the issue, which you are trying to solve.

L23: DINEOF falls from the sky here. For non-experts in the field of sea surface temperature reconstructions, it is completely unclear what this is. Also, as DINEOF appears here for the first time, it is a must that you provide a reference. The reference comes two sentences further down, which is too late. A brief description of the method would be appropriate here.

page 2: L9: "to detect the presence of non-linear, stochastic features" - I disagree that neural networks "detect" these features. Rather they are able to "learn" such features and thus potentially produce more detailed reconstructions of them.

L11: This statement is too general. I strongly suggest to first explain briefly what types of neural networks exist and how they can/could be used for the problem you want to solve (including references to the most important deep learning papers). Then you can make the argument that these networks (and in particular CNN derivates) are generally trained with complete data, whereas in your application you need to find a method which can train with scenes containing missing data, because there are no complete satellite scenes available (or only very few).

L14: this paragraph contains some of the literature review I am asking for in my previous comment. However, here it is on the one hand too specific (only ocean data applications), but on the other hand too superficial as it doesn't become clear why you need to develop a new approach and cannot simply apply for example the method of Krasnopolsky et al.

L33: I storngly suggest to re-organize the paper so that it follows the classical structure and describes the method before the data, and in particular before another reconstruction (DINEOF) is reported.

Page 3: L4: Delete the first sentence. You don't need a motivation within your "Data" section. Such an argument belongs in the introduction, if you wish to explain, for example, why you designed your study based on this dataset and not another one. In section 2 you should only describe the dataset, without any "discussion".

L19: The cross-validation deserves more explanation, because you are publishing in

a journal which is primarily read by non-experts in the field of machine learning. It is important to note that (contrary to standard image analysis) subsequent scenes from the AVHRR data are not independent. Therefore you cannot use a random sample to construct your test dataset (or "validation dataset", whichever terminology you prefer). I am wondering if 50 scenes are indeed sufficient to thoroughly test the generalization of the network. Not being an ocean scientist, I can only assume that typical transport time scales in your study region are on the order of a week(?). This would imply that the first 7 scenes of your "independent" test data are still "polluted" and thus not fully independent. Have you tried retaining a larger test sample?

L21: How can you retain > 100,000 measurements from 50 scenes? Are these individual pixels, or did you actually apply cross-validation with random sampling, thus ignoring the argument I made above?

Page 4: Figure 1: I cannot see any arrows, which are referenced in the figure caption.

L3: Again, some explanation of DINEOF is warranted in this paper. It should be clear what this method does, without having to access the referenced papers. For details you can refer to them, but not for the fundamental "explanation".

Page 5: Table 1: instead of "fewer layers" or "more layers", the number of layers should be given.

L5: I don't think this is an appropriate citation here. It is the principle of EOFs to detect relations between variables and construct an orthogonal set of linear functions to model these relations. Due to the cutoff after N EOFs, there is always smoothing applied. A proper citation here would be some standard statistics book.

L7: I am not at all surprised by this result: if you extend the timeseries, you will be more likely to sample patterns, which have not been observed before and which don't fit well to the already "learned" EOFs. Hence, there is less structure that can be described by the EOFs and more noise.

L14: I disagree that deep neural networks are "extensively" used in Earth sciences. This field is developing rapdily, but the applications are so far far from "extensive".

Page 6: L6: what does "different errors" mean? Different to what? Also: as mentioned before, for an article in this journal there needs to be a brief description of CAEs in the method section, which should contain enough information that an uninitiated reader (e.g. an ocean modeller) understands why the approach might actually work. Clearly, before the discussion taking place here, the reader must know how the network is constructed, which activation functions are used, which optimizer is used, whether regularization techniques are applied, etc. And the basis for the network built here is probably coming from some deep learning paper, which then needs to be cited. Such a description follows on page 7. Please restructure.

Page 7: L7: the references refer to convolutional layers only. However, you are applying an autoencoder approach, so the appropriate references should be made. For example: G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313(5786):504, 2006.

L9: "different numbe rof filters" - not filter sizes. Your filter size is always 3x3 as you state-of-the-art just below.

L16: composed of

Page 8: Please also write down the loss function.

Page 9: L2: So here it appears that indeed your "independent" validation data are not independent.

L9: I don't understand the random masking: is one random mask applied to each image and then the same image used in each epoch? Or do you apply different random masks to the same image as a means to augment your data and increase generalization?

L24: Remove the statement about other variables, because you focus exclusively on SST. This can go in the conclusions section, but is confusing here.

Page 10: L10: I am confused here as to how the reconstruction and the training works. Normally, you first train your network and then you reconstruct (in particular on unseen data). Then you cannot take any average between epochs 200 and 1000.

Page 13: L5: "[..] underestimate the actual error by 15% but one can argue that an underestimation of the expected error of this magnitude should be acceptable for most purposes." This sentence deserves further explanation. What is the generally accepted accuracy of the error estimate?

Page 14: L4: again: not "filter sizes" but "number of filters"

Page 15: Figure 5: the figure titles are misleading. Apparently you are always showing results for one specific day. This day should then be mentioned in the caption and nt as title on the first panel. The way the panels are labelled now suggests that you compare apples with oranges (which I don't believe you do). Also here and in Figure 6 it is not quite clear to me if the DINEOF reconstruction also had to deal with the "added clouds" or not. Higher up, when you mention the addition of random clouds it would be good to see a typical fraction of image size which is obstructed by these random clouds. From figures 5 and 6, this obstruction seems to be quite large.

Page 17: L4: where can the reader see the comparison between in-situ obs and the reconstructions? This paragraph remains qualitative and doesn't contribute anything meaningful.

L15: indeed - if the deep learning method was applied correctly (which is somewhat difficult to judge from this paper due to all the issues described in this review), then this is a very nice and important result, which shows the superiority of deep learning approaches with their ability to learn non-linear functions compared to standard statistical methods. This key result could probably be brought out even clearer.

Page 18: L2: why is this method "practical"? I assume it is, but this is only because of my background knowledge. This point needs to be made explicit somewhere in the paper. You list computation times for training, but you don't say how long it takes to reconstruct a scene once the network has been trained.

Page 19: After rewriting other parts of the manuscript I suggest to re-read the final part of the conclusions to see if the paper ends with the highest impact message.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-128, 2019.