Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-128-AC1, 2019

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.

Alexander Barth et al.

a.barth@uliege.be

Received and published: 11 December 2019

We would like to thank the reviewer for carefully reading the manuscript and providing constructive criticism. We hope that our reply below answers the questions and comments by the reviewer in an adequate manner. Our response is in bold-face.

General Comments: It is an interesting paper overall. The author uses Auto Encoder to reconstruct the missing data commonly found in optical satellite remote sensing caused by instrument failure or cloud cover. The author uses an interesting way to handle

missing data in training image. As compared to the widely used DINEOF method, the author showed that DINCAE can, on some degree, produce better results measured in RMSE metrics, as well as spatial distributions of SST. From the technique side, Auto Encoder is a commonly used machine learning method in semantic segmentation and object detection. The author uses this method to solve, particularly SST, reconstruction obtained from satellite remote sensing. It is an interesting and meaningful problem to tackle. But as for developing a new methodology, I have some concerns.

1. The applicability of this method to other satellite measurements. Variables such as SST, have low frequency variability both in space and time (If I am right). This nature suggest that it is relatively easier for CNN to estimate the spatial correlation (e.g. for an image in which there are multiple people, it is harder to do segmentation than for an image with only lawn and sky). This also gives ground that average pooling turns out to achieve better results than max pooling, as stated in paper. For variables, especially those on land, such as plant reflectance, usually have high frequency variability both in space and time due to heterogeneous growth stage, background, and so on. These feature creates additional challenges, which I think, cannot be handled with the method configuration stated in paper. It will be interesting to see how it does (This may not directly related to the topic of this paper). Additionally, the method is tested at one site,which hardly persuasive to show its applicability over the globe. Will a model trained at one site be able to use at another site, or it is needed to develop a new model to a new site, which usually needs a lot of work to prepare data, training model, parameter tuning and so on? If so, from model deployment side, what the advantage of using it?

   We think that it is quite normal to first apply a technique to a limited area before addressing the problem at global scale and we choose a parameter which is quite important for oceanography. The initial papers of DINEOF

(Beckers and Rixen, 2003; Alvera-Azcárate et al., 2005) focused also on sea surface temperature reconstruction and in subsequent papers it was shown that it can also be applied to other ocean parameters with (some adaptations) such as Chlorophyll concentration, sea surface salinity and sea surface currents. We are envisioning a similar path for the DINCAE technique. As our background is in physical oceanography, we can only speculate whether this technique can be applied to land-based data. The underlying motivation of DINEOF (and to some extent of DINCAE) is the fact that a large fraction (90% and more) of the variability of the ocean (as obtained from remote sensing data) can be explained by a reduced number modes (often 10 to 50 modes). So even when the satellite scene is partially obscured the missing data can be recovered using the data present in the satellite scene because the number of inherent degrees of freedom is relatively low. The reviewer mentions the case of plant reflectance which seems indeed to be a case where the number of degrees of freedom is apparently much higher and where it can be indeed difficult to use the same method. But we would say that this is a difficult case for any reconstruction method because the data actually measured by the satellite has less information on the data obscured by e.g. clouds. For such cases, it is particularly important to associate a reconstructed scene with a reliable error estimate. At least for sea surface temperature we were able to demonstrate that this could be done.

In the same way that the EOFs obtained from DINEOF are specific and only optimal to the studied zone, we see that the network of DINCAE is, until the contrary is proven, only specific to a given zone. The aim of this paper is to use the typical use-case of DINEOF (limited zone, ocean parameter) and to see if DINCAE can provide a better reconstruction than DINEOF. Many other reconstruction techniques used in oceanography, such as optimal

interpolation or variational analysis, require some set of parameters to be tuned to a specific site. For DINCAE it is rather the structure of the network which can be optimized for a given site but there is arguably a greater chance that these depend less of the studied site than for example parameters like the correlation length in optimal interpolation. We actually have good results using the present network structure on the Adriatic Sea and we have been contacted by a researcher using the same network architecture on the South China Sea and West Philippine Sea providing a convincing reconstruction.

For this paper we worked on a regional scale, because we believe that this matches the typical approach of oceanographic studies which focus on a specific zone of interest and improve the understanding of the processes in this area (instead of trying to understand a process directly on a global scale).

2. Temporal feature of reconstructed variables EOF method is essentially PCA analysis. DINEOF method does take into consideration of both temporal and spatial correlation of variables, to my understanding. Though DINCAE, as described in the paper, also uses the spatial and temporal correlation of variables, it only uses correlation presented in 3 days (the day, the day before and the day after). In other words, spatial information is what it uses mainly for reconstructing. Do you have persuasive arguments that 3 days correlation in time are enough to capture temporal dependency? However, longer time dependency, e.g. seasonality, may also be important on estimating missing values. In this case, network configuration both capture spatial and temporal structure of variables (e.g. LSTM + CNN) could be more general and powerful.

   The cloud cover varies normally quite rapidly from one day to the next as

it does so on the time scale of a couple of hours as revealed by geostationary satellites like SEVIRI. For polar orbiting satellite we typically have a sea-surface image every day. So the one day before and one day after, is justified by the fact that we will have a reasonable chance that a pixel covered at a given day is not covered by the day before or the day after. Providing more than just 3 days could improve the performance as it would increase the available information, but it could also increase the risk of overfitting. As a response to the reviewer, we also tried with 5 time instances but it degraded the results. The following has been added to the manuscript:

For every time instance we use the data from 3 time instances in the reconstruction: the current day, as well the data from the previous and next day. As a variant of the previous reconstruction experiment we increase the number of time instance from 3 to 5 centered at the current time instance. However, the cross-validation error for this experiment is 0.433 °C and the results are not improved. Increasing the number of input features can aggravate the potential for overfitting as the number of parameters in the neural network is increased. A combination of convolutional neural network with recurrent neural networks (like Long short-term memory, LSTM) might be a better way to include the time dependencies.

The idea using an LSTM is indeed an interesting idea. But we rather think this should be addressed in a follow-up study as we were able to show progress using the present structure of the neural network.

The present technique uses also the day-of-the-year as input of the neural network so the information about the season is available to the neural network. The day-of-the-year is transformed by a cosinus and sinus specifically to facilitate the representation of the seasonal cycle (e.g. the 1st January is as close to the 2nd January as the 31st December).

Technique Comments

Page 1 line 2: ‘A method to reconstruct missing data in satellite data using a neural network is presented’ The first sentence is not as precise as it should be. As the first impression that this paper is going to introduce a neural network based method to reconstruct/interpolate gappy satellite images caused by cloud coverage, instrument failures (e.g. LandSat 7) and so on. However the following paper mostly discussed an AutoEncoder method to reconstruct SST and tested only on SST.

While we think that the method is generic, we have only tested it on SST and thus we changed the abstract accordingly as suggested by the reviewer to make it clear that so far we only demonstrated its use for SST. The abstract now starts with:

A method to reconstruct missing data in sea surface temperature data using a neural network is presented.

This matches in fact the scope as set by the title of the manuscript which also mentioned specifically sea surface temperature.

Page 2 line 31: ‘effectively reducing....’ What is the meaning of putting this sentence here.

We think that the dimensionality reduction is a central aspect for the reconstruction of missing data. This aspect is actually shared with DINEOF. To make this clear we have expanded this paragraph.

An auto-encoder is a particular type of network which can compress and decompress the information in an input dataset (Hinton and Salakhutdinov, 2006), effectively reducing the dimensionality in the input data. Projecting the input data on a low-dimensional subspace is also the central idea of DINEOF, where it is achieved by an EOF decomposition.

Page 4. Figure 1 caption ‘The arrow represent...’ There is no arrow on figure

Unfortunately, we included an earlier version of the figure in the manuscript (without the arrows). The correct figure is the one below and the manuscript is updated.

Page 4 line 6: ‘so that for a given date also the satellite’ Delete ‘also’

Ok, done, thanks.

Page 5 line 12, ‘...in the following’ Delete ‘in the following’

Ok, done, thanks.

Page 5 line 19 ‘assimilation of data’ Change to ‘data assimilation’

Ok, done, thanks.

Page 7 line 20 ‘skip connection’ Does the resolution of SST data have effect on how you use skip connections? How large scale is called large scale for resolution of 4KM by 4KM, how about SST with resolution 1KM by 1KM. From another point of view,this

operation again consolidate to use the spatial information for reconstruction, while temporal information somehow is ignored.

For us, large-scale refers to scale of variability which affects the SST over the entire domain: for example the overall position of the main current and the heating and cooling related to the seasonal cycle. Short scale refers to mesoscale circulation features (visible also in SST) related to meanders and eddies which typically have a length-scale of 50 km.

The initial idea is that these large scales should go through the bottle-neck of the convolutional autoencoder while the small scales are handled by the skip connections (experiment labeled “DINCAE (2 skip connections)” in table 1). However, it turned out that it is beneficial to have these skip connections at all levels of the convolutional neural network (experiment labeled “DINCAE (all skip connections)”) so that a distinction between scale with and without skip connections (and large versus small scale) is no longer necessary.

Page 8 line 5 The two parameters here seemly have profound effect on reconstruction result, how does these two parameter chosen?

It is not clear to us, why the reviewer thinks that these parameters have a profound effect on the reconstruction. The range of allowed values are very far from restrictive. We added the following to clarify this point.

The effective range of the error standard deviation is thus from $exp(-\gamma/2) = 0.0067$ °C **to** $\delta^{-\frac{1}{2}} = 31.6\ °C$ **which is a relatively wide range as the error is expected to be O(0.1) to O(1) °C . The bounds are only effective during the very first epochs of the neural network where the weights are still close to random values.**

Page 9 line 15 ‘the output of the neural network is a Gaussian probability distribution’ The author assume the output is a Gaussian distribution, ‘is a Gaussian distribution’ means the author know it is Gaussian.

We agree and changed “is a Gaussian probability distribution” by “is assumed to be a Gaussian probability distribution”.

Page 10 line 18-21 ‘As mentioned before, ...neural network’ Not quite understand the training procedure here. ‘a random subset of data is marked as missing’? Since the missing data is marked randomly for each epoch, it is possible that at epoch = k, some part of data is marked as missing, while at epoch = k+1, the same part of data of the same image is marked as available. If this is the case, it essentially means the model was told what it should predict randomly? This is somewhat contradictory with Page 9 line 10.

We agree that this part is confusing and more information is added to the manuscript. First, we want to explain how a traditional auto-encoder works:

- **Some data are marked for validation and never used during training**
- **The network is given some data as input and produce an output which should be as close as possible to the input. So all training data are given at all epochs to the network**
- **The network is validated using the validation data set aside.**

So the traditional auto-enoder optimises how well the provided input data can be recovered after dimensionality reduction.

In our approach, there are two steps where data are intentionally hidden to the network:

1. **The validation data that were set aside and never used during the training (page 3, line 19 of the original manuscript), similar to the traditional auto-encoder.**
2. **Some additional data in every minibatch were set aside to compute the reconstruction error and its gradient (unlike the traditional auto-encoder). This additional subset is chosen at random.**

This is done because the main purpose of the network is to assess the ability of the network to reconstruct the missing data using the available data. In fact, we are not withholding less data than the traditional auto-encoder. The downside of the approach is that the cost function fluctuates more because it is computed only over a relatively smaller set of data. But for us this is acceptable (and controlled by taking the average of the output of the network at several epochs) because the cost function reflects more closely the objective: reconstruct missing data from the available data (instead of reproducing the input data as it is the case of the traditional auto-encoder).

The traditional auto-encoder approach trained using only clear images was not considered because only 13 images of out 5266 have a cloud coverage of less than 5%. So the ability the handle missing data was for us a requirement from the start.

Concerning the specific question “Since the missing data is marked randomly for each epoch, it is possible that at epoch = k, some part of data is marked as missing, while at epoch = k+1, the same part of data of the same image is marked as available. . . .”. The reviewer is right in its interpretation. But this is always the case in supervised learning. The gradients are computed using observations (or true labels,...) of the training dataset and observations are used multiple times

(once in every epoch). But of course, the validation dataset is never used during training and used only at the last step to assess the accuracy of the network.

Page 10 line 21-22 'we average ...intermediate result' Why do not average multiple runs?

We agree that averaging of multiple runs would be preferable but it would increase tremendously the computation time, by e.g. a factor of 30 if one would average over 30 runs for example. We added the following to the manuscript:

Alternatively one would average the output of an ensemble of neural networks initialized with different weights (and possibly using different structures) but this would significantly increase the necessary computing resources of the technique (Krizhevsky et al., 2012). But this ensemble averaging approach could be beneficial to improve the representation of the expected error and the accuracy of the reconstruction.

Page 11 Figure 2 caption 'red dash line ...' How come the average DINCAE reconstruction is smaller than RMSE at any given epoch? Also, the error curve indicates that the model has no sign of convergence. I bet if you continue training the model for another 1000 epochs, the cross validation error curves will not converge. This also indicates that there might be something wrong in the training procedure. Can you plot your lossfunction here as well?

It is quite common that the RMS error relative to a cross-validation data of a neural network does not converge. This is actually the basis of strategies like early stopping Prechelt2012. The RMSE of the average DINCAE reconstruction is smaller than the RMSE at any given epoch because computing the RMSE is a non-linear operation. The DINCAE reconstruction at a given epoch included

some variability which is not (or insufficiently) constrained by the observations. This explains also why the CV RMSE fluctuates. By taking the mean of the reconstruction at different epoch these fluctuations are averaged out and a better reconstruction is obtained. An alternative technique would be the use of an ensemble of neural networks (Krizhevsky et al., 2012) as noted also by the reviewer in his/her other comment.

The figure below shows the loss function for every minibatch. High fluctuations are quite apparent from this figure. But it is expected that the loss function using any optimization method based on mini-batch fluctuates (unless the learning explicitly is forced to zero, which is not the case here) because the loss function is evaluated using a different mini-batch at every iteration. Consequently the gradient of the cost function includes also some stochastic variability. Even if the dataset is small and the gradient could be computed over the entire dataset at once, using mini-batches is still advised because these fluctuations allow the cost function to get out of a local minima (Ge et al., 2015; Masters and Luschi, 2018). While the mini-batch selection effectively computes the gradient over a temporal subset, the additional data marked as missing within a minibatch is a spatial subset which enhances these fluctuations but allows us to define the cost function more closely to our objective (i.e. inferring the missing data from observations, as explained above). (The previous paragraph has also been added to the manuscript).

Page 14 line 16 'also tried ...' The max pooling operation tries to extract distinct signals from neighbors, while average pooling operation tries to extract common signals from neighbors. For SST, which has low frequency variation in space, it makes sense average pooling should do better than max pooling.

Thanks for this remark. We added this interpretation to the manuscript. In fact,

in the current research literature max pooling has completely replaced average pool posed in the pioneering work from LeCun et al. (1998) for CNN and image recognition. It was a surprise to see that the seemingly outdated average pooling worked better than the max pooling for our case. But we agree with the interpretation of the reviewer which has been included in the revised manuscript. Another way to look at this is the fact that for a dynamical system in the linear regime, different flow features (solution to the underlying primitive equations) coexist and contribute in an additive way to the total flow.

Page 17 line 14 '...reconstruction is it thus...' Change 'is it' to 'it is

Sorry for the typo, and thank you for reading the manuscript so carefully to the end!

References

Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers, J.-M.: Reconstruction of incomplete oceanographic data sets using Empirical Orthogonal Functions. Application to the Adriatic Sea Surface Temperature., Ocean Modelling, 9, 325–346, https://doi.org/10.1016/j.ocemod.2004.08.001, http://hdl.handle.net/2268/4296, 2005.

Beckers, J.-M. and Rixen, M.: EOF calculation and data filling from incomplete oceanographic datasets, Journal of Atmospheric and Oceanic Technology, 20, 1839–1856, https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2, 2003.

Ge, R., Huang, F., Jin, C., and Yuan, Y.: Escaping From Saddle Points - Online Stochastic Gradient for Tensor Decomposition, CoRR, abs/1503.02101, http://arxiv.org/abs/1503.02101, 2015.

Hinton, G. E. and Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks, Science, 313, 504–507, https://doi.org/10.1126/science.1127647, https://science.sciencemag.org/content/313/5786/504, 2006.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25, edited by Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., pp. 1097–1105, Curran Associates, Inc., http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf, 2012.

Masters, D. and Luschi, C.: Revisiting Small Batch Training for Deep Neural Networks, CoRR, abs/1804.07612, http://arxiv.org/abs/1804.07612, 2018.

Prechelt, L.: Early Stopping — But When?, pp. 53–67, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-35289-8_5, 2012.

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

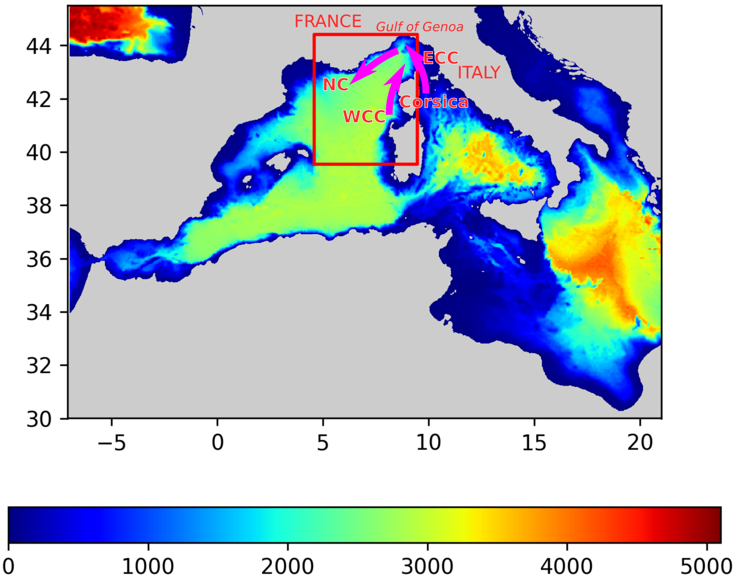

Fig. 1. The red rectangle delimits the studied region and the color represents the bathymetry in meters. The arrows represent the main currents: the Western Corsican Current (WCC), the Eastern Corsican Curren

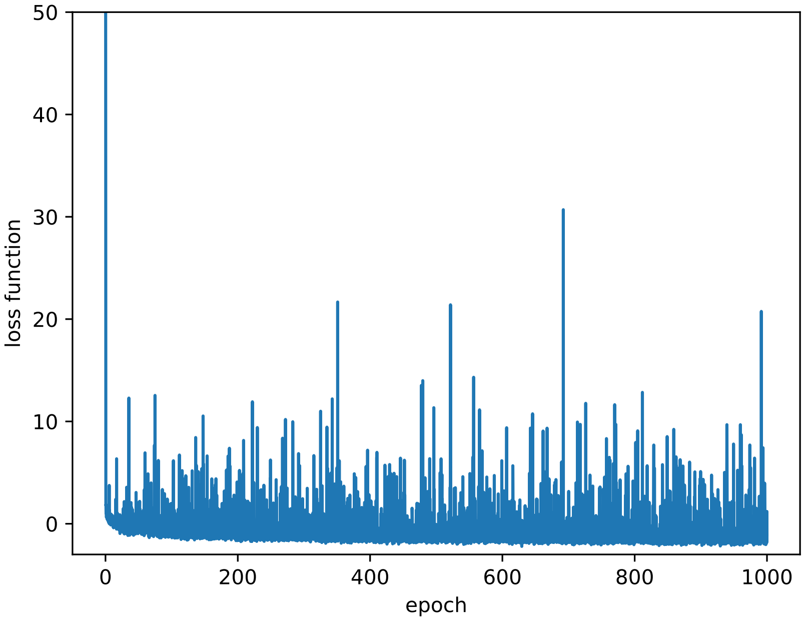

Fig. 2. The loss function computed internally for every minibatch during the optimization.