Review of the manuscript titled: Ensemble reconstruction of missing satellite data using a denoising diffusion model: application to chlorophyll *a* concentration in the Black Sea

May 2024

The manuscript titled "Ensemble reconstruction of missing satellite data using a denoising diffusion model: application to chlorophyll *a* concentration in the Black Sea" presents a novel method for the reconstruction of missing data in satellite observations. The method is based on the denoising diffusion probabilistic model, where a sample form a noise distribution is incrementally transformed to a sample from the target distribution. In this case, the authors use this approach to fill-in missing patches in the satellite observations, with the aim of: creating a spatially coherent reconstruction, and to create an ensemble of reconstructions, which better describe data variations compared to a single mean estimator for the missing value. At each step in the noise reduction process, the authors use a U-net style neural network to predict the necessary noise delta, which, if applied to the the current step input, reduces the total amount of noise in the image, approaching the target reconstruction. The authors test their method on the task of reconstructing chlorophyll *a* values for a select region over the Black Sea, with good results implying, that the method embodies a meaningful addition to the roster of spatial data reconstruction techniques.

1 General comments

Overall, the manuscript is structured well and clearly demonstrates the application of denoising diffusion probabilistic models in the domain of satellite reconstruction, with compelling results when compared to the baseline method, denoted as DINCAE (a method which was previously applied on this task). However, before I can fully recommend the manuscript for publication, there are some shortcomings that have to be addressed.

Firstly, there are some implementation details which might indicate potential errors in the algorithm's implementation. It seems that the presented equations pertaining to the diffusion model somewhat deviate from the standard definition, while the changes lack an explanation or motivation.

Secondly, the results section, although demonstrating that the proposed method compares favourably to the baseline approach in terms of RMSE, variogram, and quality of reconstruction, could still benefit from some additional comparisons. Furthermore, I do not wholly agree with some of the conclusions reached by the authors with regards to the interpretation of the Talagrand diagram.

I provide detailed arguments for each of the issues raised in the following Section.

2 Specific comments

2.1 Denoising diffusion probabilistic model implementation

- Equation (3): The authors state that the conditional distribution of the image x at step t given x_0 in the forward pass is defined as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, \bar{\alpha}_t\boldsymbol{I})$. This suggests that the value of the variance approaches zero as t increases, reducing the distribution to be zero mean and zero standard deviation in the limit. The value of the variance in the conditional case, if I am not mistaken, should be equal to $(1 - \bar{\alpha}_t)\boldsymbol{I}$, given the transformation defined by Equation (1).

- Equation (9): The reverse probability $p_\Theta(x_T)$ should not be parameterised by Θ, since the distribution is defined in Equation (5), where no such parameters are present. If these distributions indeed differ the authors should explain what properties the parameterization defines in this specific case. Additionally, a technical mistake seems to be present in the term $q(x_{1:T})|x_0$, which should be equal to $q(x_{1:T}|x_0)$ correct?

2.2 Diffusion model neural network description

This comment pertains to the neural network description provided in paragraphs 185 and 190, and Table 1. The definition of the neural network is given recursively, with each block l being dependent on the block $l-1$. However, given how the levels are provided $l = 1, ..., L$ (L being the depth level) and $C_l = [16, 32, 64, 128]$, this description might be confusing for readers unfamiliar with the architecture. For example, one can make the mistake that the number of channels on the first level C_1 is equal to 16. However, as far as I understand the provided description, the block at depth level 4 contains 16 channels while the block at depth level 1 contains 128 channels. Therefore, the initial block corresponds to $l = 4$, while the "deepest" block corresponds to $l = 1$, which seem counter-intuitive given that l denotes the depth level.

Consider these two cases: if $l = 1$ and $C_1 = 16$, then the inner block of block $l = 1$ is an identity and the recursion stops immediately. However, if one assumes that $l = 4$ and $C_4 = 16$, then the inner block at $l = 3$ contains $C_3 = 64$ etc. which results in the familiar U-net architecture, where the spatial dimension is reduced with each consequent block and the number of channels increases. This later assumption is not self evident from the provided description. Therefore, I suggest that the author either flip the depth indices l, such that $l = L, ..., 1$, or that they flip the C_l values $C_l = [128, 64, 32, 16]$ while keeping the indices intact.

2.3 Description of the DINCAE method

The authors provide a short description of the DINCAE method's training setup in the Results section, in paragraphs 250 and 255. While I believe that this is beneficial to the manuscript the description somewhat breaks the flow of the Results section. Therefore, I suggest that the authors move this description to the Appendix.

2.4 Talagrand diagram in Figure 10

The authors compute the Talagrand diagram using the ensemble as an approximate distribution function, where each ensemble memeber represents an equally probable event realization. The authors sort the ensemble members in an ascending order for each masked pixel, independently. The resulting empirical distribution functions and their corresponding ground truth values are used to construct the diagram. The histograms for both the *dev* and *test* datasets are displayed in Figure 10 and the authors conclude that: "*Figure 10 shows the Talagrand diagram computed for the development and testing dataset. It can be seen that except for the two first and two last bins (corresponding to the probabilities between* 3% *and* 97%*), the Talagrand diagram is relatively flat. This shows that the produced probabilities are reliable, except for very rare events where the produced ensemble is underdispersive. The difficulty of predicting rare events is a known issue in machine learning (e.g. Kaiser et al., 2017) and a dedicated area of research.*"

Here I would like to raise a minor concern regarding the use of "*probabilities are reliable*" in this context. The produced probabilities are *marginally reliable*, since each pixel is treated independently from its neighbours. However, this does not necessarily imply that the joint distribution of the ensemble is reliable, which is not decernable from the conclusion reached by the authors. For example, consider taking the same ensemble forecast produced in this work, however, with its values randomly permuted between the corresponding members for each pixel. A permuted forecast like this would exhibit the same Talagrand diagram (since the sorting on a per-pixel basis restores the initial diagram conditions), however, the forecast would not be jointly reliable as the spatial relationships would be lost. Therefore, I suggest that the authors state that this evaluation method, as is, assesses the marginal reliability only and not the joint. Again, this is not a major concern for readers familiar with the evaluation technique, however, since the spatial correlation is an important asset of this proposed reconstruction method, a clarification of this would be welcome.

However, I do not agree that the excess number of observations in the extreme ranks implies that the method perform poorly for very rare events only. The excess denotes that the distribution described by the ensemble exhibits short tails, meaning, that a disproportionate number of observations fall into those ranks. These observations can including realizations that are not rare at all and should actually be described by other ranks. Therefore, I would suggest rewording this conclusion such that it reflect the notion of the distribution tails being too short rather than an explicit comment on the reliability of rare event forecasting.

2.5 DINCAE comparisons

The output of the DINCAE method can be interpreted as a normal distribution, where the reconstruction is its mean, and the reconstruction error its standard deviation (or variance), correct? If so, I suggest constructing the Talagrand diagram for the DINCAE method as well, which would further demonstrate the impact of the proposed method's distributional capabilities compared to the baseline. The same comment applies to the evaluation using the CRPS method, where the DINCAE approach (given that the above assumption holds) can also be included.

I also suggest that the authors include the training and inference times for the DINCAE method, as well as the number of parameters of the DINCAE method, such that the reader can better asses the relative computational complexity of this new approach compared to the baseline.

2.6 Diffusion model performance conditional on the number of valid input image pixels

The proposed diffusion model is dependent on the valid pixel (pixels without clouds) in the input image to construct a spatially consistent reconstruction. This approach produces realistic reconstructions with a high degree of spatial correlation as can be seen in the provided examples. This, however, prompted the following consideration: how does the performance of the reconstruction degrade in relation to the number of valid pixel available in the input image? An evaluation like this could be an interesting inclusion in the current manuscript, providing a practitioner with the knowledge on how reliable the reconstruction is given how much information is present in the original input image. The ensemble spread might already describe such notions however, it might not be marginally reliable when considering images with a high degree of missing valid data.

2.7 Miscellaneous comments

- On "*For the validation and test data, we randomly took the cloud mask from other time instances to mask additional grid cells which will be used for validation. Only images with a cloud mask between 15% and 35% of the missing date were considered as an additional mask to obtain a sufficient number of "clouded" pixels without masking an image almost entirely.*" in paragraph 150: Does this mean that, when constructing an input image from the validation/testing datasets, a random image with 15% to 30% of missing data is selected (still from the same dataset) and its cloud mask is used to cover the current input image's pixels? If so, it seems that this approach can still result in a completely covered image if the image being masked has a coverage greater than 85%, correct? Or are only images with a coverage of less than 70% considered for the validation/test datasets? A few comments on this would be appreciated.

- On "*During training, for each image of the training dataset, a different image is randomly selected (also from the training dataset) and its cloud mask is used to degrade clear pixels of the input image (Figure 3). The stage of degradation t of these pixels is randomly chosen between 1 and T.*" in paragraph 170: Can it not occur that the training image can be fully degraded after the additional cloud mask is provided (example: input image has 20% valid data and the mask has a cover of 80%)? Such training images might slow the convergence of the method as the denoising process is completely unguided. Or is this event rare in practice?

  Furthermore, what is the benefit of setting the degradation value between 1 and T instead of just T? If I understand correctly, during inference, each missing pixel is treated as being fully degraded. Is there

a difference in performance compared to setting all pixels to the fully degraded value T? Does this help in cases where the training image might be degraded to a high spatial degree (above example)? A few comments on this would be appreciated.

- On "*For each image of the validation and test two datasets, the reconstruction process is repeated 64 times, leading to an ensemble of possible reconstructed fields.*" in paragraph 230: How did you choose the number of ensemble members (64 members) in the reconstruction? Was it determined empirically? If so, please provide an explanation.
- On "*In Barth et al. (2020), it has been shown that the accuracy of a reconstruction can be improved by averaging the obtained reconstruction over a certain number of epochs after the epoch 200.*" in paragraph 250: I do not fully understand this approach. Does this mean that, during training, intermediate models from epoch 200 onwards are saved and the mean reconstruction from each of those models is used as the final DINCAE output?

3 Technical comments

- Figure 2, Figure 6, Figure 7, Figure 8: Consider adding lat, lon labels to the axis.
- Broken Latex mathematical symbol for $\bar{\alpha}_T$ in paragraph 180.
- The kernel size k_s (Table 1) does not require a subscript since it is a fixed value across levels. Consider omitting the subscript.
- "As an illusion" in paragraph 215: misplaced use of the word "illusion". Consider replacing with "illustration".
- Table 2: Typo in "desactivated". Additionally, consider explaining the meaning of the rows of the table as some are not self evident, for example "refinement step".
- On "*(corresponding to the probabilities between 3% and 97%)*": Should this not be equal to "between 1.5% and 98.5%" since each interval has a weight of $\frac{1}{65}$ implying, that the first rank covers realizations with a probability of occurrence between 0 and 0.015 and the last rank between 0.98 and 1? Therefore, the middle ranks exhibit a coverage between 0.015 and 0.98, correct?