1 Reviewer 2

1.1 General Comments

> This research work proposes a new method that can be applied in the reconstruction of missing data in geophysical datasets, and more specifically cloudy satellite images. The method is applied successfully to chlorophyll images, improving the skill of state of the art reconstruction methods. The paper is well written, scientifically consistent and contains enough novel contributions. Without doubt it deserves publication, but after some technical clarifications/corrections and a couple of small extensions of the results. Thinking in a broad audience dealing with the problem of missing data in geophysical datasets, some parts of the document might not be very reader friendly, and look biased towards the computer vision community. Considering the scope of this journal, an effort in that direction would be appreciated. The relatively small area considered is the main weakness of the study and introduces some concern about the applicability of the technique elsewhere. However, as the diffusion model is already trained using the complete Black Sea, it would be possible to extend the analysis by repeating the analysis (with the same hyper-parameters) for other areas (dynamically similar and not) to produce a figure like Figure 9 for multiple locations (see details later).
>
> Identification of specific comments and technical corrections: P==page; L==line

We are grateful for the careful reading and the detailed comments from Reviewer 2. We have addressed the comments from the reviewer in the revised manuscript. We agree that the small area is a weakness of our study. In the revised manuscript, we applied this technique to 9 additional domains of the Black Sea and compared the results of the diffusion model with the result of DINCAE. For each of these domains we also computed the variogram (Figure 9 of the original manuscript) as the reviewer suggested. We found that, overall, the conclusion is robust when applying this technique to other areas of the Black Sea. More information is in the detailed reply below.

1.2 Specific Comments

> P2L41: DINCAE refers to version in Barth et al., 2020; or alternatively to modified version in Barth et al., 2022? To both?

In fact, this statement refers to both versions. We added the reference to the 2020 and 2022 papers to clarify this point. Thank you for pointing this out.

> P3L67-73: for the broad audience dealing with geophysical missing data reconstruction techniques, this could be made a bit less technical and more descriptive

We have expanded this section in the hope that this description is now more easily understood by the typical geophysical audience of this journal. The revised paragraph now reads:

> In the classifier-free guidance algorithm (Ho and Salimans, 2022), the neural network denoising the images also depends explicitly on the class label. While training the neural network, this class label is sometimes replaced by a null label (i.e. a vector with all elements equal to zero). As a result the trained neural network can either denoise *any* image of the training dataset (when given the null label) or a specific subset of the training dataset (matching the provided label). During sampling the reverse diffusion is steered by a scaled difference between the noise predicted knowing the label and the noise predicted with a null label and therefore enhancing the similarity of the generated image with the provided label.

P3L86-P4L90: it is not clear if the mean and std are these of a pixel over the group of images, or alternatively are calculated image by image

We have normalized the images by using the mean and standard deviation of the entire training dataset. We clarified this in the revised manuscript.

P4L91-93: alpha and beta parameters are identified as diffusion step t dependent. True? Make that more explicit if it is the case. For the ease of understanding also make explicit that degradation level increases with increasing t

Yes, α and β depend on the diffusion step. We modified the revised manuscript at different places to make it more explicit.

P4L106: implications for this application of "small steps sizes beta" not very clear

We added that this has to be understood in the limit where the discrete diffusion process tends to the continuous diffusion.

P6L135: image frequency is? 3h, 6h, daily...?

We indeed forgot to mention that this is a daily dataset. This is corrected in the revised manuscript.

P6L140: why 20%? Conservative approach? Sensitivity detected for certain threshold?

This is our first implementation of a denoising diffusion model so we adopted a conservative approach here. To our knowledge it is also the first implementation of a denoising diffusion model trained on incomplete satellite data. Standard diffusion models thus require 100 % of valid data. One should also consider that during training we mask additional points using the cloud mask of other images and that the loss function considers only these additional masked pixels (where the ground truth) is available. If we would use a lower threshold, say 10%, and apply an additional mask (with potentially up to 90% of missing data), there would be several instances of image without any pixels that can be considered in the loss function. This could significantly increase the training and with unclear benefits. Related to this question, is the remark from the other reviewer how does the network behave if there is a significant amount of missing data. We have grouped the results into ranges of cloud fraction and the highest is 85%-95%.

P6L138-143: training and validation datasets are defined here, but later also test dataset is referred.

and

P15L248-250: previous comments on the distinction of training, verification and test periods apply here in comparison with development phase mentioned now

It was actually defined on line 141 in the original manuscript:

> The **validation** and **test** data range from September 1, 2021 to August 31, 2022 and from September 1, 2022 to August 31, 2023, respectively. [emphasis added]

In the revised manuscript, we splitted the sentence in two for clarity:

> The validation dataset is composed of the 12 months of data between September 1, 2021 to August 31, 2022. The following 12 months (from September 1, 2022 to August 31, 2023) are used as test data.

P6L140: is that the number of training images or the number after breaking the original figures in tiles?

Yes, we added this is the revised manuscript:

> Only tiles with at least 20% valid data (i.e. non-clouded pixels) are used for training to reduce training time. In total, there are 851926 images (after splitting the data into tiles) for training.

P6L139: reason for the 64x64 tile splitting is given later; here is confusing without justification; say choice is justified later?

We agree and changed the order in this section. In the revised manuscript we start the relevant paragraph with the justification.

P6L146: what is meant with DINCAE being only applied with a fixed location? Does it mean that while the diffusion model is trained using data over the complete area DINCAE is applied only to the small box in Figure 2, and hence they are only compared over that small location?

Yes, this is correct. While we could have trained DINCAE also with data from other locations, DINCAE has only been tested and validated so far when trained with a fixed location. We preferred to keep the implementation of DINCAE as it was discussed in the previous published papers. The diffusion model could not be trained using a fixed location as discussed on lines 216 - 219 of the original manuscript.

P6L146-147: that justifies the use of a small area for testing purposes, but why not extend the comparison to other locations in the area (by for example adding extra small areas of the same size, maybe also with some overlapping?)
and
P15L248: regarding proposal of extension to other areas made in the “general comments”, this would imply training of DINCAE for such areas in this step.

In the revised manuscript we applied the diffusion and DINCAE to addition regions in the Black Sea. For overlapping domains, one can compute ensemble statistics such as mean and variances provided by the diffusion model which should be consistent from one domain to the next overlapping domain. However, this is not the case of the individual ensemble members which are currently independently. We plan to address this limitation in future follow-up work.

The following as been added to the manuscript:

Further domains are considered to test the applicability of the trained diffusion model in comparison with DINCAE to explore the different dynamical regimes. In Figure 13, the domain used previously is labeled as 1, and the additional domains are labeled 2 to 10. For each of these domains DINCAE is trained using only the data from the corresponding domain using the hyperparameters presented in Table B1. As the diffusion model is trained using 64 x 64 tiles from the whole Black Sea, it is not trained again but used only in the inference mode. The RMS error for each domain is shown in table 1 and the corresponding variogram can be seen in Figure 14. Overall the results from the previous test on the first domain are also applicable to other domains. The RMS error of the diffusion model is lower than the corresponding RMS error of DINCAE except for domain 7. At the same time, the variance for all domains across different scales is more realistic for the diffusion model.

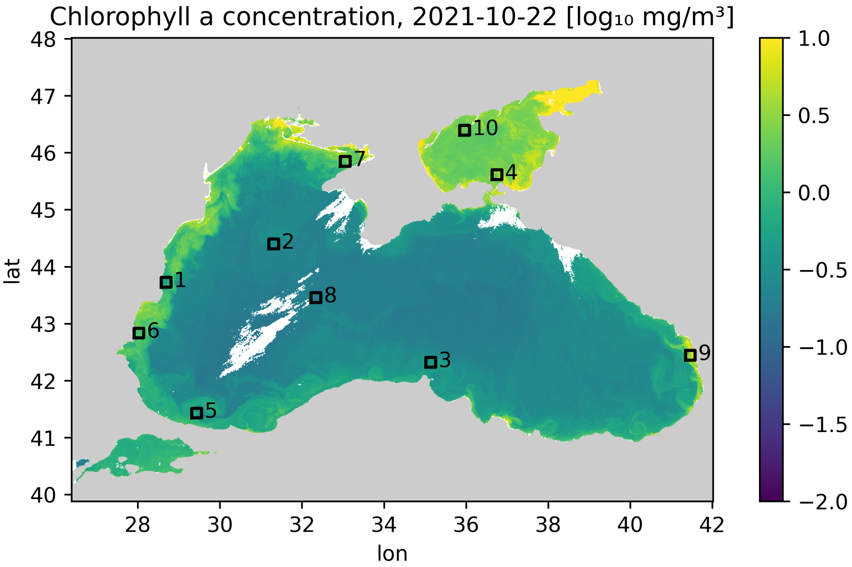

Figure 13: Additional domains where the diffusion model is applied (domain 2 to domain 10)

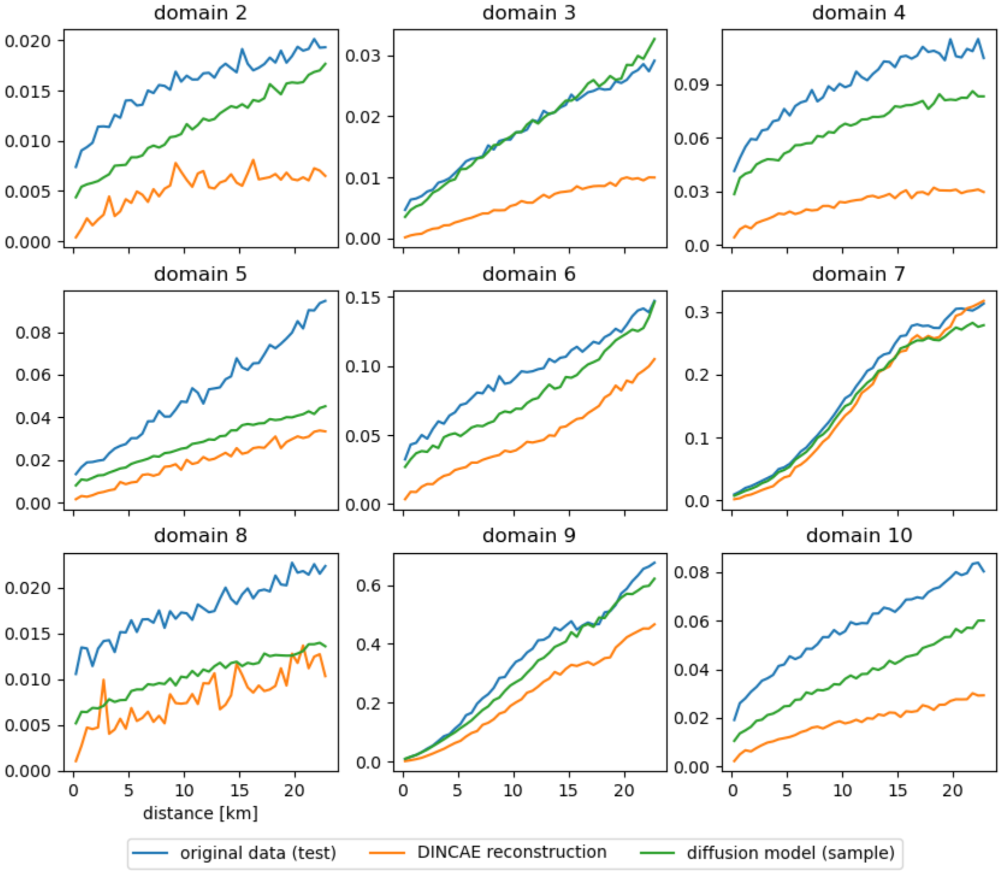

Figure 14: Variogram for the independent test data

Table 1: RMS error relative to the independent test data for different domains.

domain	RMS DINCAE	RMS Diffusion Model	std(obs)
1	0.175	0.163	0.331
2	0.159	0.058	0.226
3	0.225	0.056	0.211
4	0.162	0.155	0.253
5	0.162	0.074	0.251
6	0.182	0.143	0.353
7	0.090	0.096	0.295
8	0.119	0.062	0.286
9	0.189	0.149	0.442
10	0.116	0.111	0.244
median	0.158	0.107	0.289

P6L146-149: why is this relevant? Why proceed this way? It seems relevant for the image reconstruction community, but is that the case for an audience dealing with geophysical data reconstruction?

This sentence has been removed in the revised manuscript.

P9L165: eq. 1 also?

Yes, equation 1 should also be mentioned here. The manuscript is updated. Thank you

P9L165: how does it ensure spatial coherence?

To clarify this point we added the following to the paragraph in question:

> It is important to note that all operations in the training and sampling algorithms (equations 1, 10 and 11) are only pointwise operations (i.e., operations that apply to each grid cell independently) that do not involve the neighboring grid cells, except for the neural network which ensures spatial coherence. The spatial coherence is mainly due to the convolutional layers whose weights have been trained to provide the same spatial structure as in the training dataset.

P9L173-174: it refers to the pixels of the added cloud mask? T is randomly selected for each pixel in that mask or the whole mask? From figure 4 and explanation in P11L196-197 it looks that is shared for all pixels in the mask...

Yes, this is correct. The step T is shared among all additional masked pixels. This has been clarified in the text:

> The stage of degradation t of these pixels is randomly chosen between 1 and T but applied uniformly to all withheld pixels. This is important because the noise is reduced progressively during inference and the neural network needs to know how to handle intermediate degradation levels.

P9L177 and Figure 4 caption: "Scaled diffusion step" of figure lacks explanation at this point (comes later); a basic description in the figure caption would help

We agree and extended the caption with the information: "The diffusion step t ($0 \leq t \leq T$) is scaled linearly to the interval $-\frac{1}{2}$ and $\frac{1}{2}$"

P10L180-182 and Figure 4: the way "predicted noise" in figure 4 is created from the "partially corrupted image", based on the description within this line, is not easy to understand and could be more explicit (how the neural network operates to create the figure)

The output of the neural network is a 2D field aiming to predict the noise that has been added to the input. The neural network can predict the added noise because it learned the typical spatial structures in the training dataset and it is able to recognise them even in a corrupted image. At a first approximation, the neural network acts like a high-pass filter to identify the noise, which is then removed iteratively during sampling.

> P10 Figure 5: this figure is not mentioned in the text and it is not clear how it interacts with the neural network; if it is part of it or a separated process…

A reference to this figure has been added by addressing the following issue here below.

> P11L198 would indicate that the training operates in the forward direction, while P5L127-129 that it is the same trained network that is applied, but in the reverse mode, to produce a ensemble of possible reconstructed versions… True? Make this clearer

During training, noise is intentionally added to the image (advancing from diffusion step l to $l+1$) and the neural network is trained to predict the noise, allowing it to denoise the image and go from step $l+1$ back to l.

The following has been changed in the sampling section of the revised manuscript:

> After training the neural network, the missing data in the validation and test dataset are reconstructed. Every clear pixel of the input image is considered to be in the non-degraded state $t = 0$ and all other pixels (clouded or on land) are in the fully degraded state $t = T$ and initialized with normal distributed random values. For the later pixels, the reverse diffusion process is used iteratively (going from step $l + 1$ to l) to reduce their noise keeping the originally present pixels unchanged (Figure 5).

> P11L197: then -1/2 refers to initial or non degraded while 1/2 is fully degraded? Make that explicit

This is correct. We added this information in the revised manuscript.

> P11L205-212: No comment about future or present public code availability?

The source code for training and inference of the neural network is now available at the address: `https://github.com/gher-uliege/DINDiff.jl`.

> P11L211-212: training and validation datasets were presented in “Data” section; does this refer to “validation” dataset? Is there a third “test” dataset?

As mentioned in another comment above, the test data were actually defined on line 141 (data section) in the original manuscript. In the revised manuscript, we introduced the test and validation data in two separate sentences for clarity.

P12L222: same as previous comments

As mentioned before, the paragraph has been revised:

> After training the neural network, the missing data in the validation and test dataset are reconstructed. Every clear pixel of the input image is considered to be in the non-degraded state $t = 0$ and all other pixels (clouded or on land) are in the fully degraded state $t = T$ and initialized with normal distributed random values. For the later pixels, the reverse diffusion process is used iteratively (going from step $l + 1$ to l) to reduce their noise keeping the originally present pixels unchanged (Figure 5).

We hope that this is now clearer.

P12L225: comment about "P9L165" on the spatial coherence made before clarifies here, maybe is enough to mention that only here

We added specifically that the convolutional layers of the U-Net ensure spatial coherence in the method section.

P13L229: any reason to select that particular ensemble size?

We chose 64 ensemble members as a compromise between the diversity of ensemble members and computational time and guided by the fact that the ECMWF real-time S2S forecasts use 51 members for its ensemble forecast. We updated the manuscript accordingly.

P13L241: Blue for zero std? Mask out using another out of the color bar (white,...)?

A standard deviation of zero for initially present pixels is indeed intended. All ensemble members will be identical where the pixels are initially present. As a result the corresponding standard deviation is zero. The ensemble members are only different where the initial pixel is clouded. The area in uniform blue is thus not a mask and indeed a true zero.

We expanded the paragraph in the following way in the revised manuscript:

> For every ensemble member, the interpolated fields in the pixels for which we have valid values in the input data is, per construction, identical to the initial input value. The ensemble standard deviation at these locations is thus consequently equal to zero.

P15L257-258: RMS error of the diffusion model is calculated as the ensemble average of individual RMS values?

In fact, the RMS error of the diffusion model is based on the ensemble mean. We added this information to the revised manuscript.

P15L259-260 and Tables 3 and 4: It would be interesting to add some metric about the intra-ensemble variability of the RMS for the diffusion model (in Tables 3 and 4): RMS_max and RMS_min and/or variance/std of the RMS inside the ensemble. This would provide a basic idea about the quality of individual reconstructions inside the ensemble.

As suggested, we compute the RMS error for every ensemble member individually and compute the minimum and maximum of these error statistics. It is important to note that the averaging operation for the RMS runs over all space and time dimensions (but obviously considering only pixels where there are actual measurements but whose value has been withheld). The revised paragraph now reads:

> In all cases, the bias is relatively low and does not contribute significantly to the RMS error. The RMS error of the diffusion model (based on the ensemble mean) is slightly smaller than the RMS error of DINCAE for development and test datasets. However, as expected the RMS error of every ensemble member individually is substantially larger than the RMS error of the ensemble mean. Given that the RMS error is computed over all time instances, the RMS error for a single ensemble member is relatively stable. The maximum and minimum RMS error among the 64 ensemble members are 0.202 and 0.211 $\log_{10}$ mg m^{-3} respectively.

P18L270: "randomly chosen locations" including a certain number of comparisons? all possible?

We agree that this was not clear enough. We added the following part:

> Here we are considering a variogram only as a function of distance $h = \|\mathbf{x}_1 - \mathbf{x}_2\|$, which allows us to estimate the variogram numerically by computing the squared differences for the field at randomly chosen locations. These squared differences are averaged over bins of distances using all time instances of the validation and test datasets. As many different random locations were chosen until there are at least 10000 pairs for each bin of distance.

P18L271-272: that means that comparisons are made for individual members and then averaged over the ensemble to produce the metric value?

Yes, this is correct. Following has been added to the revised manuscript to clarify this:

> For the diffusion model, the variogram is deduced using the individual ensemble members, and the averaging in equation (12) is done also over different ensemble members.

P18L278: "does not converge to zero" for the original data and the diffusion model, not for DINCAE?

This is correct. We added the following to the text:

> DINCAE effectively removes (or significantly reduces) the spatially uncorrelated white noise and therefore the corresponding variogram shows a clear tendency towards zero for smaller distances.

P20L295-297: meaning of the over estimated upper and lower limits?

The corresponding paragraph has been rewritten and now reads:

> ... It can be seen that the error statistics of the diffusion model are closer to the ideal flat curve for the diffusion model than for DINCAE. This shows that the probabilities produced by the diffusion model are marginally reliable, except for the tails of the marginal PDF (first and last bin, corresponding to the probabilities between 1.5% and 98.5%) where the produced ensemble is underdispersive.

P20-21L299-320: Although interesting, the analysis of the CRPS seem somehow weak without any reference to compare with...

The other reviewer made a similar comment and asked to produce the relevant scores for DINCAE. While DINCAE is not able to produce a full ensemble (or full pdf), the CRPS score relies only on the marginal pdf, which can be provided by DINCAE. The corresponding Talagrand diagrams have also been produced. The table with the CRPS results and the accompanying discussion has been extended in the revised manuscript.

P22L336-342: multivariate reconstructions are mentioned, but what about univariate reconstructions like for Sea Surface Temperature or Sea Surface Salinity, for instance, how is the method expected to behave?

We are optimistic that the method can also be applied to sea surface temperature, as the spatial and temporal scales in these images are more coherent and less patchy than those in chlorophyll data, which should help in the reconstruction. However, for sea surface salinity having large area (almost) constantly masked due to radio frequency interference could pose a problem for the diffusion model as it is not clear if the neural network can learn the appropriate scales for these regions.

1.3 Technical corrections

P2L33-35: the reference to the error of initially missing values is confusing

This sentence has been rephrased as:

> For satellite images where all missing data have been reconstructed, it is clear that the error of the reconstructed and initial missing pixels is typically larger than the error of the original pixels.

P2L51-53: hard to understand, could be clearer? An example maybe?

We have added the following example in the hope that this is now clearer:

> Since multiple images would be consistent with the partial information present, a neural network trained to minimize e.g. the mean square error, would then implicitly produce the average of all these possible states. For example, if the exact position of a front is not visible in a satellite image, a reconstructed image would have the tendency to smooth out the front as it is implicitly the average of multiple images with the front in different positions. Consequently, this means that small scale information cannot be adequately retained.

References

Ho, J. and Salimans, T.: Classifier-Free Diffusion Guidance, https://doi.org/10.48550/arXiv.2207.12598, 2022.