Dear Editor and Reviewers,

We sincerely thank both reviewers for their careful reading of our manuscript and for their constructive comments. We are pleased that the potential of the proposed diffusion-based framework for sea surface salinity downscaling was recognized, while we also acknowledge that several aspects of the original manuscript required clarification and a more cautious interpretation.

In the revised manuscript, we have substantially clarified the scope of the study. We now present the work explicitly as a controlled GLORYS-based proof-of-concept experiment, designed to evaluate the methodological potential of guided diffusion models under idealized conditions. We have revised the abstract, introduction, method section and discussion to avoid overclaiming regarding immediate satellite applicability, and the superiority of the proposed method over deterministic baselines. We have also added a dedicated discussion of the challenges associated with the transfer to real satellite SSS observations, including SSH and SST products availability, satellite SSS uncertainties, validation with sparse in-situ observations, and uncertainty estimation from the diffusion ensemble.

Below, we provide a point-by-point response to the reviewers' comments.

Response to Reviewer 1

Comment:

1. Use of high-resolution SSH: The manuscript relies on high-resolution SSH information. However, prior to the launch of SWOT, SSH observations were primarily derived from conventional altimetry, which has a spatial resolution comparable to that of SSS (approximately 0.25°). This limitation should be explicitly acknowledged. The authors should clarify whether the intended application relies on SWOT observations, which would inevitably constrain the temporal coverage of the downscaled SSS products.

We agree with the reviewer. In the original manuscript, the use of high-resolution SSH was not sufficiently discussed in relation to the actual availability of satellite altimetry observations. We have revised the discussion to clarify that, before SWOT, conventional altimetry products generally provide information at mesoscale resolution and do not offer the same fine-scale coverage as the 1/12° SSH used in our GLORYS-based experiment.

We now present the high-resolution SSH configuration as a controlled experiment designed to evaluate the potential information content of a dynamical auxiliary variable, rather than as a directly available operational configuration for the full historical satellite SSS period. We also discuss that future use of SWOT-derived SSH could provide finer-scale constraints, but would necessarily limit the temporal coverage of any future downscaled product to the SWOT era. We therefore now make clear that the SSH-conditioned experiment should be interpreted as methodological, not yet operational.

Comment:
2. Use of high-resolution SST: The use of high-resolution SST is strongly affected by cloud coverage. SST can be retrieved from microwave sensors, which provide observations at resolutions comparable to SSS, or from infrared sensors, which offer much higher spatial resolution but only under cloud-free conditions. This limitation should be discussed. In particular, the authors could highlight that, under cloudy conditions, the SSH-only version of the algorithm may become the preferred alternative, thereby emphasizing the added value of that approach.

We first emphasize that the resolution of GLORYS is 1/12° and is used for learning the joint distribution of the three variables SSS, SST, SSH (this corresponds to the forward phase). This allows the model to learn a GLORYS-consistent multivariate prior over SSS, SST and SSH. The resolution of the variables used for the backward phase could be less precise (they are used to estimate in the estimated numerical model the more probable triplet of images (SSS,SST,SSH). This is an important point for the transfer of the method to real satellite observations. In the present GLORYS-based proof-of-concept, SST is available everywhere at 1/12° resolution and therefore provides an idealized high-resolution auxiliary constraint. This situation differs from real satellite SST products. Infrared SST observations can provide kilometre- to few-kilometre-scale information, but only under cloud-free conditions. Under cloudy conditions, the high-resolution infrared constraint is missing, and Level-4 SST products rely on multi-sensor merging and interpolation, often combining coarser microwave SST, previous or nearby clear-sky observations, in-situ measurements and the assumptions of the analysis method. In the present paper we simplify the problem and we assume that only SSS is degraded and that the SST and SSH are known at 1/12°. The backward phase shows how we can reconstruct the SSS 1/12° knowing the 2 other variables. The same backward procedure can be used with a degraded version of the three variables simultaneously. This is not done in the present paper limited to the theoretical aspect of diffusion models.

We have clarified this point in the revised discussion. Although products such as SST-CCI, OSTIA or MUR may be distributed as daily gap-free fields on 0.05° or finer grids, their local effective information content can degrade under persistent cloud cover. In such cases, the SST field should not be interpreted as a uniformly reliable high-resolution observation, but rather as an auxiliary constraint with spatially variable uncertainty and effective resolution.

From the methodological point of view, guided diffusion models can in principle handle imperfect or noisy conditioning information: the reconstruction is not imposed solely by SST, but results from a balance between the learned SSS prior, the coarse SSS constraint, and the auxiliary SST/SSH fields. Therefore, degraded SST information does not necessarily make the method unusable, but it may reduce the quality of the reconstruction where the SST constraint is weak, interpolated or uncertain. This should be accounted for in future satellite applications, for example by using uncertainty-aware weighting of the auxiliary constraints.

Comment:
3. Impact of SSS uncertainties: Errors in SSS observations, particularly in coastal regions, are a critical issue. These errors are not limited to random noise but often include systematic biases and other retrieval artifacts. It is therefore important to

discuss how sensitive the proposed methodology is to such errors. The manuscript would benefit from a sensitivity analysis evaluating the impact of different noise levels and error structures on the reconstruction.

We fully agree that this is a critical issue for any future application to real satellite SSS observations. In the original manuscript, the coarse SSS input was treated as a synthetically degraded version of the GLORYS reference. This is useful for a controlled proof-of-concept experiment, but it does not reproduce the full complexity of real satellite Level-4 SSS products.

We have revised the manuscript to make this limitation clearer. In particular, we now discuss the ESA CCI SSS product as an example of the type of satellite product that could be considered in future applications. CCI SSS is a multi-sensor Level-4 product derived from L-band satellite missions such as SMOS, SMAP and Aquarius. Although it is sampled on a 0.25° grid, its effective spatial resolution is approximately 50 km, reflecting the native radiometer footprints and the Level-4 merging procedure. Therefore, it should not be interpreted as a simple clean low-pass version of the true SSS field.

We now also clarify that CCI SSS provides uncertainty and quality information that would be essential for any future observation-guided diffusion approach. These include the random error field, the percentage of unexplained SSS variability, the number of observations, the number of rejected outliers, and quality flags for general SSS quality, land-sea contamination and ice-sea contamination. Such information could be used to weight the observation constraint spatially, reduce the influence of uncertain or contaminated pixels, and mask unreliable regions. The observation operator used in this paper is intentionally simple and should be replaced, in future satellite applications, by a product-aware operator including the effective satellite footprint, missing-data masks, uncertainty fields and quality-control flags. In particular, products such as ESA CCI SSS provide per-pixel uncertainty and quality information that could be used to spatially weight the observation guidance, rather than imposing the same constraint everywhere.

Preliminary developments outside the scope of this manuscript with real CCI SSS suggest that uncertainty-aware guidance will be necessary when the input SSS contains heterogeneous errors. Because these experiments involve a different observation operator, satellite-specific uncertainty weighting and independent in-situ validation, we consider them as a separate extension of the present proof-of-concept rather than as part of the current manuscript. We now mention this perspective more clearly and identify a systematic sensitivity analysis to random, correlated and biased errors as a required step before operational satellite application.

Comment:
4. Training strategy using observations: An important aspect that should be discussed is how the model would be trained using real observations. The most widely used reference salinity dataset within the satellite community is provided by the Argo network. However, Argo observations do not provide high-resolution spatial coverage and are particularly sparse in coastal regions, which are among the most relevant application areas of the proposed method. The authors should discuss how this limitation could affect the training and validation of the algorithm.

Diffusion models cannot be used with a limited amount of data, in order to learn a joint distribution of images, as for Deep Learning algorithms, they need a coherent “big” dataset. The objective of the present framework is not to train the model directly from Argo observations. Argo profiles provide accurate pointwise salinity measurements, but they do not provide spatially complete, temporally coherent, high-resolution two-dimensional SSS fields. They therefore do not contain the spatial and temporal organization needed to learn the fine-scale physical structures targeted by the downscaling problem.

In the present proof-of-concept, the role of GLORYS is to provide a physically consistent gridded reference from which the diffusion model can learn a high-resolution prior. More precisely, the forward learning stage learns the statistical and dynamical coherence represented in GLORYS, including the spatial structures of SSS and their relationships with SST and SSH. In this sense, the model learns what a physically plausible high-resolution triplet of SSS, SST and SSH looks like within the GLORYS framework.

The observational constraint is introduced at inference time, not during the training of the prior. In the present manuscript, this constraint is a synthetic coarse SSS field generated from GLORYS through a known degradation operator. In a future satellite application, the constraint would instead be a real satellite SSS product, such as CCI SSS, SMOS or SMAP Level-4 SSS. The reconstruction problem would then consist of finding a high-resolution SSS field that remains coherent with the learned GLORYS prior while satisfying the available satellite SSS, SST and SSH constraints through a product-aware observation operators.

This perspective is conceptually close to data assimilation: the learned diffusion prior plays a role similar to a dynamical/statistical background, while the satellite SSS product provides the observational constraint. The final reconstruction is not obtained by directly learning from sparse in-situ data, but by searching for a physically plausible high-resolution state compatible with the observations. Within this framework, Argo has a different and essential role: it provides independent validation, calibration and bias assessment. Argo observations can be used to evaluate whether the reconstructed SSS field improves upon the original satellite product and/or the reanalysis background at independent matchup locations.

Comment:
5. Uncertainty estimation and ensemble information: More generally, I believe the manuscript does not fully exploit the capabilities of the proposed methodology. Since the method generates an ensemble of possible reconstructions, the spread among ensemble members, e.g. their standard deviation, could potentially be used as an estimate of reconstruction uncertainty. Such information would be extremely valuable to the community. In addition, it is not entirely clear why the selected output corresponds to the realization closest to the ensemble mean rather than the ensemble mean itself. The rationale behind this choice should be explained.

For a given coarse SSS constraint, the stochastic inference can generate several plausible high-resolution reconstructions from different random initializations. This was illustrated in Fig. 4, where several generated SSS patches are shown for the same input, together with a pixel-wise ensemble variance map. However, we now clarify more explicitly how this spread should be interpreted.

In the revised manuscript, we describe the ensemble spread as a diagnostic of reconstruction ambiguity. Where the ensemble spread is weak, the generated members converge toward similar high-resolution structures, suggesting that the reconstruction is well constrained by the coarse SSS observation and by the auxiliary SST/SSH information. Where the spread is larger, several fine-scale configurations remain compatible with the same constraints. These regions often correspond to fronts, filaments or dynamically active structures, where small spatial displacements can produce larger differences between ensemble members.

We have also clarified the distinction between ensemble spread and formal uncertainty. The variance map shown in Fig. 4 is not presented as a fully calibrated uncertainty estimate. It is a first confidence indicator derived from the conditional ensemble. We now state this explicitly to avoid over-interpreting the ensemble variance.

Regarding the selection of the final reconstruction, we clarified why we use the realization closest to the ensemble mean rather than the ensemble mean itself. The ensemble mean is useful and may reduce pixel-wise error, but it can also smooth fronts and filaments when different plausible realizations contain slightly displaced structures. Moreover the mean realization does not correspond to a GLORYS triplet, because the problem could be multivalued. The closest-to-mean realization is used as a representative ensemble member: it remains close to the center of the conditional distribution while preserving the spatial texture of an individual physically plausible realization seen within GLORYS.

Comment:
L92–96: Please explain the impact of the normalization procedure, zero mean and unit variance, on the results. The manuscript also states that this approach reduces computational cost. Could the authors clarify this point? Computational cost typically depends on matrix dimensions, yet the proposed approach appears to increase the number of matrices being processed. Is the method parallelized? Additionally, does the dimension reduction have any impact on the accuracy of the final reconstruction?

Response:
We thank the reviewer for pointing out this unclear wording. We have revised this section. The zero-mean and unit-variance normalization is applied independently to each variable using statistics from the training period. Its purpose is to place SSS, SST and SSH on comparable numerical scales, stabilize optimization, and prevent variables with larger numerical ranges from dominating the training loss. It does not, by itself, reduce the dimensionality of the problem or the computational cost.

The computational saving comes from the patch-based training strategy, not from the normalization. Training on 64×64 patches reduces the memory cost of each training sample compared with full-domain images, while increasing the number of training examples available to the model. We have corrected the wording accordingly. We also agree that patches and ensemble members can be processed in parallel, but that the main computational advantage comes from patching rather than from normalization.

Comment:
Figure 1: Consider including the standard deviation of SSS at the low resolution used

in the right plot of the figure. This would help illustrate the loss of variability associated with the loss of resolution.

Response:

We thank the reviewer for this suggestion; we added to figure 1 the standard deviation of the degraded resolution.

Comment:
L70 and Figure 2: Have the authors considered using Sea Level Anomaly, SLA, instead of SSH? SLA generally captures clearer information on mesoscale eddies and related structures.

Response:
We thank the reviewer for this useful suggestion. In the present study, we used SSH because it is directly available in the GLORYS fields used to build the controlled experiment and because our goal was to test the information content of the available surface dynamical variable. However, we agree that SLA could provide a more focused representation of mesoscale eddies and anomaly structures by removing part of the large-scale background.

We have added this point to the discussion and now identify SLA as an important alternative conditioning variable to test in future work. We also acknowledge that SLA may be more appropriate than absolute SSH for satellite-based applications, especially when the objective is to constrain eddy-related salinity structures.

Comment:
L108: Please specify the masking procedure that is applied.

We thank the reviewer for pointing out this ambiguity. The wording in the original manuscript was inaccurate. In the experiments reported in the present manuscript, no masking procedure and no additional noise were applied to the coarse SSS observations. The coarse SSS field was generated only by applying the spatial averaging and subsampling degradation operator to the high-resolution GLORYS reference field.

Noise and masking experiments were considered separately during the development of the method, but they are not included in the present manuscript because they require a dedicated analysis of error structure, missing-data geometry and uncertainty-aware guidance. We chose to focus this paper on the fully completed controlled experiment, in order to describe the methodology and evaluate its potential under well-defined conditions.

We have therefore corrected the manuscript by removing the statement suggesting that noise or masking was applied in the reported experiments. The revised text now explicitly states that the current results correspond to a clean synthetically degraded coarse SSS observation.

Comment:
L111: Replace “both” with “both together”.

Response:
Corrected.

Comment:
L197: What is meant by an “MSE-type algorithm”? Please clarify.

Response:
We agree that this wording was imprecise. We have replaced “MSE-type algorithm” with a clearer explanation. The revised text now states that deterministic models trained with mean-square-error losses tend to favor conditional averages, which can lead to smoother reconstructions and weaker gradients when multiple fine-scale solutions are compatible with the same coarse observation.

Comment:
L310: The reported differences appear relatively small and are of similar magnitude to the improvements claimed for other methods earlier in the manuscript.

Response:
The reviewer is correct that the differences in RMSE between RESAC and the standard diffusion configurations are relatively small. We have therefore revised the interpretation of this comparison to avoid presenting the diffusion model as providing a large improvement in classical pixel-wise metrics.

RESAC remains a strong deterministic baseline. Its good performance is expected, since it was specifically designed for downscaling and relies on a different theoretical framework, with a multi-resolution control of the reconstruction that is partly motivated by scale-invariance and fractal considerations. If the comparison is restricted to RMSE or SSIM, RESAC and DIFF-SST-SSH should therefore be considered broadly comparable in the present controlled experiment.

The added value of the diffusion framework is different. It does not only provide a single deterministic reconstruction, but samples possible high-resolution fields that are consistent with both the learned prior and the imposed observations. In that sense, the guided diffusion formulation is closer to an inverse-problem or data-assimilation perspective: the learned GLORYS prior defines physically plausible high-resolution SSS/SST/SSH structures, while the coarse SSS observation constrains the reconstruction at inference time.

This formulation makes it possible to generate an ensemble of plausible reconstructions, to estimate reconstruction ambiguity from the ensemble spread, and to adapt the guidance strategy at inference time without retraining the full model. These aspects are not captured by RMSE alone and are not available in the same way in a deterministic reconstruction such as RESAC.

We have therefore revised the manuscript to state that the standard diffusion model does not strongly outperform RESAC in RMSE, but provides complementary capabilities: probabilistic sampling, ensemble-based confidence diagnostics, and flexible observation-guided reconstruction.

Comment:
L313 and Figure 7: Figure 7 does not clearly demonstrate that “DIFF-SST-SSH-GE sharpens fronts and filaments relative to DIFF-SST-SSH and RESAC.” In the pdf file the four plots look almost identical.

Response:
We agree with the reviewer. We have softened this statement and revised the interpretation of Figure 7. We no longer state that the sharpening is clearly demonstrated by the visual

comparison alone. Instead, we now describe Figure 7 as a qualitative illustration and rely more on quantitative diagnostics, such as gradient distributions and spectral analyses, to assess the structural effect of the GE procedure.

Comment:
Figure 8, L320–L325: While the figure shows improvements in some parts of the gradient distribution, there is also a significant negative difference, indicating that DIFF-SST-SSH-GE captures fewer pixels with weak gradients than observed in the reference distribution. Could this be interpreted as evidence of excessive sharpening? This negative difference is not discussed in the manuscript and deserves further explanation.

Response:
We agree with this interpretation. The negative difference at weak gradients indicates that DIFF-SST-SSH-GE reduces the proportion of weak-gradient pixels relative to the reference, which can indeed be interpreted as a sign of excessive sharpening in some parts of the distribution.

We have added this discussion to the revised manuscript. We now present GE as a trade-off rather than as a uniformly better reconstruction. The GE procedure can increase gradient contrast and partially compensate for smoothing, but it may also overcorrect the gradient distribution by converting too many weak gradients into stronger gradients. This limitation is now explicitly discussed.

Comment:
Figure 9, L325–L332: Figure 9 provides additional evidence that DIFF-SST-SSH-GE may overestimate small-scale variability. At nearly all temporal scales, the amplitude of the reconstructed signal exceeds that of the reference field. This point should be discussed more thoroughly.

Response:

This interpretation is relevant and we have expanded the discussion of Figure 9 accordingly. We now discuss this result together with the gradient-distribution analysis in Figure 8. Both diagnostics point to the same trade-off: GE increases structural contrast and reduces some of the smoothing affecting the standard reconstruction, but it can also over-sharpen the field by reducing the proportion of weak-gradient pixels and increasing the energy at higher temporal frequencies. GE should therefore not be interpreted as a uniformly superior reconstruction, but as an optional contrast-enhancing inference strategy whose strength must be tuned according to the target application.

Comment:
General comment: It would be useful to provide a clearer discussion of the added value of RESAC relative to DIFF-SST-SSH.

Response:
We agree. We have expanded the comparison between RESAC and DIFF-SST-SSH. RESAC remains a strong deterministic convolutional baseline and performs similarly to the diffusion model under classical pixelwise metrics. We now acknowledge this more clearly.

The revised manuscript emphasizes that the added value of DIFF-SST-SSH is not a large improvement in RMSE, but rather the generative and flexible nature of the framework. Unlike a deterministic model, the diffusion model can sample multiple plausible high-resolution fields, provide ensemble spread diagnostics, and modify the conditioning or guidance strategy at inference time. This distinction is now made clearer in the discussion.

The guided diffusion formulation shares with variational assimilation the idea of combining a prior representation of the system with observational constraints during reconstruction.

Comment:
L336: Please justify the statement: "their gradient magnitudes remain, on average, weaker than those of the reference fields by roughly 10%." It is not clear how this estimate was obtained from the presented results.

Response:

We thank the reviewer for pointing this out. The statement was intended as a qualitative indication of the residual underestimation of gradient magnitudes, but the numerical value of 10% was not explicitly derived or documented in the presented results. To avoid unsupported quantification, we have removed this sentence from the revised manuscript.

Response to Reviewer 2

Major comments

Comment:
As a reanalysis product incorporating data assimilation, how does the GLORYS dataset ensure dynamical consistency among SST, SSH, and SSS? Given that these variables are assimilated independently from diverse observational streams, the physical coherence between them may be compromised.

Response:
This is an important point, and we have clarified the role of GLORYS in the revised manuscript. We do not assume that GLORYS is a perfect representation of the true ocean, nor that data assimilation guarantees perfect physical consistency between SSS, SST and SSH at all scales. GLORYS combines an ocean model with heterogeneous observational constraints, and local inconsistencies may remain, especially in regions where the assimilated observations constrain the different variables unevenly.

However, in the present proof-of-concept, GLORYS is used as a controlled gridded reference from which the model learns a joint high-resolution prior. The diffusion model is trained on the joint distribution of SSS, SST and SSH fields represented in GLORYS. In other words, the model does not learn SSS independently from the other variables, but learns the spatial organization and cross-variable relationships between the three fields as they are represented in the reanalysis. The generated high-resolution states retrieved during the backward phase therefore remain constrained by this learned GLORYS-consistent multivariate distribution.

This is the reason why we view the guided inference step as conceptually close to an inverse-problem or data-assimilation framework. The learned diffusion prior provides a reduced statistical representation of physically plausible SSS/SST/SSH triplets in the GLORYS framework, while the coarse SSS observation constrains the generated state through the observation operator. The reconstruction is thus not only a pixel-wise interpolation of SSS, but a search for a high-resolution state that is both compatible with the imposed SSS constraint and coherent with the multivariate structures learned from GLORYS.

We have also clarified an important limitation. The model used in this manuscript does not explicitly learn temporal dynamics: time is represented only through the ensemble of daily GLORYS fields used for training. Therefore, the learned prior includes the range of spatial structures and cross-variable relationships sampled over time, but it does not enforce temporal continuity or dynamical evolution between consecutive days. A natural extension would be to develop a spatio-temporal diffusion model, in which several consecutive time steps are used jointly so that the temporal evolution of fronts, eddies and filaments is also learned and constrained.

Comment:
The horizontal resolution of the GLORYS dataset, 1/12°, is insufficient to resolve submesoscale and small-scale structures. Resolving submesoscale features typically requires resolutions of 1–2 km, while small-scale processes demand resolutions of hundreds of meters or higher. Consequently, it is questionable whether the proposed method can genuinely recover these fine-scale dynamics from such coarse input data.

Response:
We agree with the reviewer. The original manuscript used the terms “submesoscale” and “fine-scale” too broadly. A 1/12° reanalysis cannot fully resolve true submesoscale dynamics, and the proposed method cannot recover physical processes that are absent from the reference dataset. Nevertheless, the same methodological framework could in principle be applied to higher-resolution numerical simulations that better resolve submesoscale processes, for example kilometre-scale regional or basin-scale configurations such as NEMO NATL60/eNATL60.

We have revised the manuscript throughout to avoid overclaiming. We now state that the method reconstructs fine-scale frontal and filamentary variability at the 1/12° GLORYS reference resolution, rather than true submesoscale dynamics. The objective of the present study is therefore to recover the structures present in the high-resolution reference field from a degraded coarse observation, not to generate unresolved dynamics below the effective resolution of GLORYS.

Comment:
As shown in Table 2, the DIFF-SST-SSH-BEST configuration yields only a marginal 7% RMSE reduction compared to RESAC, while the DIFF-SST-SSH-GF variant performs worse, exhibiting a 12% higher RMSE. These inconsistent results raise concerns regarding the robustness and added value of the proposed methodology.

Consequently, the model performance requires substantial improvement to justify its novelty.

Response:
We agree that the improvement over RESAC is modest and that the original manuscript should not have overstated it. We have revised the discussion accordingly. The standard diffusion configuration is now presented as comparable to RESAC in pixelwise accuracy, rather than clearly superior.

We also clarify the interpretation of the different diffusion configurations. The BEST configuration has been renamed or explicitly described as an oracle diagnostic, because it selects the realization closest to the known reference field and is therefore only possible in a synthetic experiment. It is not an operational reconstruction. The GE configuration is now presented as a structural enhancement variant that can improve some gradient or spectral diagnostics but may degrade RMSE and SSIM. We therefore now describe the diffusion framework as offering a trade-off between pixelwise accuracy, structural realism, and probabilistic sampling, rather than as a uniformly better deterministic downscaling method.

Comment:
In general, the unknown variable should be Y, the known variable should be X. It should be redefined for equation (1) to illustrate the downscaling clearly.

Response:
We understand the reviewer's point. In many supervised learning formulations, X is often used for the known input and Y for the target variable. However, in this manuscript we adopted the notation commonly used in inverse problems and data assimilation, where X denotes the unknown high-resolution state to be estimated, Y denotes the available observation, and H is the observation operator that maps the state space into the observation space.

In our case, X is the unknown high-resolution SSS field on the 1/12° grid, while Y is the known coarse SSS observation on the 1/3° grid. The observation operator H represents the degradation from the high-resolution state to the coarse observation, through spatial averaging and subsampling. The downscaling problem is therefore to estimate X from Y, using auxiliary variables such as SST and SSH.

We have clarified this notation in the revised manuscript to avoid ambiguity.

Comment:
Both the forward diffusion process and the denoising process are described in section 3.2. It is contradictory to the title.

Response:

Fixed, we thank the reviewer for noting out this inconsistency.

Minor comments

Comment:
Line84: how can obtain a 1/3° SSS field by applying a 3×3 spatial averaging from 1/12° native field? Does it be 4×4?

Response:
We agree. This was an inconsistency in the manuscript. Since 1/3° corresponds to four times 1/12° in each horizontal direction, the degradation from 1/12° to 1/3° should be described as a 4×4 averaging and subsampling operation, if the target resolution is exactly 1/3°. We have corrected the text and checked that the degradation operator is consistently described throughout the manuscript.

Comment:
Line 114 in equation (2): the prediction should be high-resolution or low-resolution field? Lots of variables should be checked again to make sure which one is low resolution and which one is high resolution in Section 3 Method. It is difficult to understand now.

Response:
The prediction in Eq. (2) is indeed the high-resolution SSS field. This was already the convention used in the manuscript: (X) always denotes a field on the 1/12 degree grid, while (Y) denotes the corresponding coarse observation on the ⅓ degree grid. Therefore, (\hat{X}) is also defined on the 1/12 degree grid.

We understand, however, that this convention was not repeated clearly enough around Eq. (2), which may have made the distinction between low-resolution observations and high-resolution reconstructions difficult to follow. We have therefore added a short clarification in Section 3 to explicitly state the resolution associated with each variable.

In this notation, (H) is the observation operator that maps the high-resolution field (X) from the 1/12 grid to the low-resolution observation space of (Y) on the 1/3 grid. The downscaling problem is therefore to estimate the high-resolution state (X), or (\hat{X}), from the known low-resolution observation (Y), using the auxiliary high-resolution fields (Z).

Comment:
Line 129: delete the second we.

Response:
Corrected.

Comment:
Line 228: what is “the actual desired output images”? It is not clear how to compute the DIFF-SST-SSH-BEST.

Response:
The wording “actual desired output images” was indeed imprecise. What we meant is the known high-resolution GLORYS reference field available in this synthetic experiment. We have replaced this expression by “reference high-resolution GLORYS fields” or “reference high-resolution SSS field”, depending on the context.

We also clarified the meaning of DIFF-SST-SSH-BEST. In the diffusion framework, the model does not simply learn a deterministic mapping from a coarse SSS image to a high-resolution SSS image. It learns a reduced multivariate representation of the ocean state, here based on two or three physically related surface variables, such as SSS, SST and SSH. During inference, different random initializations generate different high-resolution realizations that remain compatible with the learned GLORYS multivariate prior and with the imposed coarse SSS constraint.

DIFF-SST-SSH-BEST is therefore not an additional operational model. It is an oracle diagnostic computed only in the controlled GLORYS experiment. For each test case, we generate an ensemble of high-resolution candidates using the same trained DIFF-SST-SSH prior. We then select the realization that is closest to the known high-resolution GLORYS reference, using the RMSE criterion. This selection requires access to the reference field and therefore cannot be used when downscaling real satellite observations, where the true high-resolution SSS field is unknown.

The purpose of DIFF-SST-SSH-BEST is to estimate the potential headroom of the generative prior: it indicates whether the ensemble generated by the diffusion model contains a realization closer to the reference than the representative member selected without access to the truth. It should therefore be interpreted as a theoretical lower-bound or oracle diagnostic, not as a deployable reconstruction strategy.

Comment:
Table 1: Why the SST and SSH of 1/12° are output?

Response:
This point revealed an ambiguity in the wording of Table 1. The scientific target of the downscaling experiment is the reconstruction of high-resolution SSS. SST and SSH are not evaluated as final downscaled products in the same way as SSS.

However, in the diffusion configurations, the model is trained on multichannel GLORYS fields and therefore learns a reduced representation of the ocean surface state involving SSS together with SST and/or SSH. During inference, the reverse diffusion process generates a coherent multichannel high-resolution sample. Depending on the configuration, this sample may include SSS/SST, SSS/SSH, or SSS/SST/SSH channels. These associated SST and SSH channels help the generated state remain within the learned multivariate GLORYS-consistent distribution, while the evaluated downscaling output remains SSS.

We agree that SST and SSH could alternatively be viewed only as auxiliary inputs in a purely SSS-focused proof-of-concept. However, generating them as part of the multichannel state keeps the same observation-operator logic for all variables. This is useful for future extensions in which SST or SSH may also be incomplete, degraded, or only partially observed. For example, in a more operational setting, high-resolution SST may be unavailable under cloud cover or may come from near-real-time products with spatially variable uncertainty. In that case, the same framework could condition the multichannel generated state using partial observations of SSS, SST and/or SSH through variable-specific observation operators.

We have clarified Table 1 and the associated text to avoid suggesting that SST and SSH are final products of the present study. They are generated channels of the multivariate diffusion state, whereas SSS remains the only variable evaluated as the downscaling target.