Reviewer #1:

This study presents a novel conditional flow-matching method for SST gap-filling. The authors provide a clear exposition of the method, extensive evaluations, ablation studies for various design choices, and the comparison against prior state-of-the-art ML approaches suggests this method has strong potential for operational applications. This study is of high quality and novelty and the extensive evaluations and ablations make this a particularly strong contribution. With some minor revisions (see below) it is suitable for publication in GMD.

Congratulations on a great paper and I look forward to seeing it published in GMD.

Best wishes,

Scott A. Martin

Response:

We thank the reviewer for the constructive comments and the encouraging assessment. We address each point below, indicating where the manuscript was changed and quoting the revised text.

Comment 1: The proposed method re-injects the raw observations back in pixels where they are available. While this ensures small error against available observations it is not necessarily desirable. Observations contain sensor noise and errors themselves, which under this approach will propagate into the reconstructed SST fields. Secondly, this approach will induce strong artifacts in each ensemble member reconstruction at the border between observed and unobserved regions, complicating downstream scientific analysis and interpretation. I think this limitation is just inherent to the approach chosen in this study, so I'm not suggesting the authors modify the method, but the manuscript would benefit from a discussion of this limitation and the motivation for this design choice somewhere in the text.

Response: *We agree that this trade-off deserves explicit discussion, and we have added it to the new Discussion section. We note that the residuals combine reconstruction and observation error, that FUSE is a deliberate choice made for hard observation consistency, and that the model does not depend critically on it, as the ablation without observation re-injection (Sec. 4.3.1) still reaches state-of-the-art accuracy. On the second point, we did not observe boundary discontinuities in our reconstructions, which remain spatially coherent across the observed/reconstructed transition (Figure 11c). The relevant passages now read:*

> *Discussion (new section):*
>
> *"DIRECT reconstructs an L3 satellite product, which is itself a retrieval carrying sensor noise and residual cloud contamination. Our metrics are computed against this product, so the reported residuals combine reconstruction error with the observation error already present in the reference. The FUSE step re-injects the measured pixels at every integration step and thus inherits their noise, a deliberate trade-off in exchange for hard observation consistency. The model does not depend*

critically on this choice: the variant that disables observation re-injection (Sec. 4.3.1) still reconstructs the missing regions to state-of-the-art accuracy, showing that performance is driven by the generative prior rather than the re-injection alone."

Section 4.2.4 - Power spectrum density analysis

"[...] For CRITER, however, this small-scale energy is largely non-physical: the gradient-magnitude fields (Figs.10c and 11}c) reveal grid-like block artifacts, a byproduct of patch-based processing. In contrast, DIRECT's energy profile reflects coherent, albeit slightly smoothed, oceanic structures. The same gradient-magnitude fields show no discontinuity tracing the boundary between observed and reconstructed regions (Fig 11c), indicating that the FUSE re-injection does not introduce a visible seam despite anchoring the output to measured pixels at every step. DIRECT therefore provides a more physically faithful representation of the continuum, avoiding both the spurious grid-like artefacts that inflate the PSD of patch-based baselines and the boundary artifacts that might be expected from observation re-injection."

Comment 2: The SST reconstruction methods benchmarked against here (as in the CRITER paper) are all experimental ML methods. While this helps to establish DIRECT as state of the art among ML approaches, the reader is left unclear on whether any of these methods outperform non-ML approaches like objective analysis used in real operational L4 products. The manuscript would be significantly strengthened by including one or more of the operational L4 products (e.g. from CMEMS or PO.DAAC) in the evaluation to the extent possible.

Response: *Following this suggestion, we added a comparison against operational gap-free L4 ocean temperature products from the Copernicus Marine Service: reprocessed OSTIA sea surface temperature and the ESA SST CCI and C3S analyses, produced by variational data assimilation and optimal interpolation rather than machine learning. The results and their interpretation are given in a new subsection 4.2.1 and Table 2. Since the L4 products carry a systematic offset relative to our L3 reference, we estimate a per-day mean bias over the observed region, subtract it from each L4 field, and report the resulting centred RMSE. Even after this correction, DIRECT matches the L3 reference considerably more closely in all regions, indicating that the remaining gap is structural rather than a mean offset. This is a measure of agreement with our L3 reference rather than a like-for-like skill comparison, since the L4 products also ingest overlapping sensors and in situ data (OSTIA). For convenience, the new subsection is reproduced below:*

Section 4.2.1(Comparison with operational L4 products)

"We compare DIRECT against operational gap-free Level-4 (L4) SST products, based on variational data assimilation and optimal interpolation, distributed by the Copernicus Marine Service: the reprocessed OSTIA analysis (E.U. Copernicus Marine Service Information (CMEMS), 2026b), the ESA SST CCI and C3S

reprocessed analyses (E.U. Copernicus Marine Service Information (CMEMS), 2026a). Each product was co-located with SST from our three domain datasets and evaluated as a candidate reconstruction against the same L3 reference used throughout. This comparison should be read as a measure of agreement with our L3 reference, not of reconstruction accuracy. The L4 products also ingest in situ data and data from the same sensors as our L3 source, meaning the pixels we synthetically withhold were in general available to the L4 products. Their error over these pixels is therefore not an out-of-sample gap-filling error as it is for DIRECT. The L4 products additionally carry a systematic offset relative to our L3 reference (Table 2, bias column); to prevent this offset from dominating the comparison, we estimate a per-day mean bias over the observed region, subtract it from each L4 field, and report the resulting centred RMSE (CRMSE).

Table 2 shows that, even after this bias removal, DIRECT remains closer to the L3 reference than any of the operational products in all three regions. Relative to the best-performing L4 product in each region, $CRMSE_{mis}$ is lower by 44 % on the Mediterranean, 37 % on the Adriatic, and 8 % on the Atlantic, indicating that the remaining gap reflects a difference in spatial structure rather than a mean offset. Note that this does not imply that DIRECT is correspondingly more accurate than objective analysis as a technique; it reflects that DIRECT is trained to reconstruct this specific high-resolution L3 product, whereas the L4 analyses target foundation SST and cross-sensor, temporally consistent fields rather than fidelity to any one regional L3 source. A fair skill comparison would evaluate both against independent in-situ observations such as Argo profiles, which neither assimilates, a direction we leave to future work."

DIRECT vs L4 products:

DATASET	Model / Product	CRMSE all [˚C]	CRMSE mis [˚C]	CRMSE obs [˚C]	Bias [˚C]
Mediterranean	DIRECT	**0,113**	**0,234**	**0,000**	0.000
	ESA CCI	0,401	0,421	0,393	0,239
	C3S	0,399	0,420	0,392	0,283
	OSTIA	0,408	0,429	0,400	0,241
Adriatic	DIRECT	**0,113**	**0,208**	**0,001**	0.000
	ESA CCI	0,301	0,329	0,290	0,186
	C3S	0,324	0,355	0,313	0,175
	OSTIA	0,348	0,378	0,336	0,119
Atlantic	DIRECT	**0,363**	**0,489**	**0,001**	0.000
	ESA CCI	0,533	0,542	0,520	0,235

	C3S	0,522	0,531	0,510	0,275
	OSTIA	0,540	0,550	0,528	0,356

Comment 3: DIRECT leverages only the sparse, high-resolution L3 SST observations but microwave sensors provide additional observations that penetrate clouds but with coarser spatial resolution. It would be interesting to consider how the DIRECT approach could be modified in future to incorporate these valuable mesoscale-resolving constraints beneath clouds. Not suggesting the authors modify the method for this study, but it would be an interesting future research direction and could be discussed in the Discussion section.

Response: *We were honestly intrigued enough by the suggestion to implement a simple DIRECT extension using the microwave data and added it to the ablation results – since we treat this as another test of injecting conditioning to our flow model. Concretely, we added an ablation (Section 4.3.8, Table 13) in which a coarse-resolution microwave OI product is supplied as an additional input through a dedicated branch. The effect on accuracy is marginal, thus we report the experiment as a demonstration of extensibility rather than an accuracy gain, and discuss the fuller treatment as future work in the Discussion. For convenience, we reproduce the added ablation and the related discussion below:*

> *Section 4.3.8 (Incorporating microwave observations):*
>
> *"DIRECT operates on high-resolution infrared L3 observations, which are unavailable under cloud cover. Microwave radiometers, however, do retrieve SST through clouds, albeit at substantially coarser spatial resolution, and therefore offer a potential source of information beneath clouds. Because DIRECT ingests its spatio-temporal context through a lightweight input projection, additional observation sources can be added with minimal architectural change. To demonstrate this, we conducted a preliminary experiment on the Mediterranean dataset using the daily microwave optimally-interpolated SST product MW OI SST v5.1(Remote Sensing Systems, 2022), regridded and co-located to our domain. The microwave field is supplied as an additional input by embedding it through a separate convolution branch and summing it with the projected context, mirroring the treatment of C_{2t} (denoted $DIRECT_{MW}$).*
>
> *As shown in Table 13, the microwave variant leaves reconstruction accuracy essentially unchanged, with $RMSE_{mis}$ marginally lower than the default model. We attribute the small effect to the coarse resolution of the microwave product relative to the infrared L3 fields. The purpose of the experiment is not to demonstrate an accuracy gain but to show that DIRECT is readily extensible to heterogeneous observation sources. A more careful treatment of the resolution mismatch and observation-specific uncertainty, together with an evaluation in regimes of persistent, large-scale cloud cover where cloud-penetrating observations are expected to be most valuable, is a promising direction that we leave to future work."*

Discussion:

"The temporal context uses a fixed three-day window, following prior evidence that most relevant temporal information lies within a few surrounding days. This is limited when informative observations fall outside the window, most notably during cloud systems that persist over many days. Cloud-penetrating microwave observations are well suited to these conditions, and our preliminary experiment (Sec. 4.3.8) shows that DIRECT incorporates such a source with minimal architectural change, making a more thorough integration targeted at persistent-cloud regimes a natural next step."

Dataset	Model	RMSE all	RMSE mis	RMSE obs
Mediterranean	DIRECT	0,113	0,234	0,000
	MW Branch	0,112	0,230	0,001

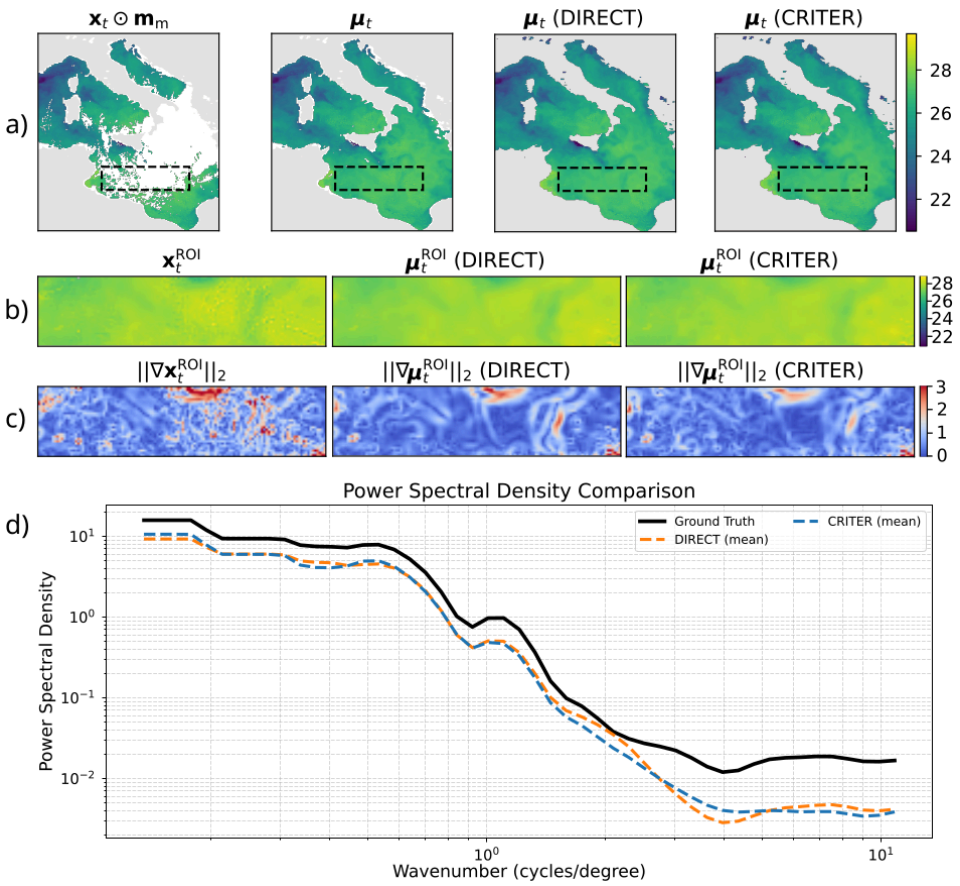
Comment 4: Figs 10 & 11 present isotropic PSD of SST. My understanding is these were computed by taking 2D spectrum of total SST field and then azimuthally averaging? There appears to be a large-scale anisotropic SST background in the ROI which will distort the inferred isotropic PSD. I think this may be part of why all the curves lie so closely on top of each other at all but the highest wavenumbers. A more appropriate metric would be to compute these spectra on the SST anomaly from a coarse background (e.g. linear plane fit to the ROI?).

Response: *This is an excellent idea. As the reviewer suggested, we now subtract a least-squares linear plane fitted to each field over the ROI before windowing and computing the FFT, so the PSD is computed on the anomaly relative to this background. Figures 10 and 11, their captions, and the text in Section 4.2.4 are updated (we also relabelled the sub-panels a–d). The conclusion is unchanged. The revised passage reads:*

Section 4.2.4 (Power spectral density analysis):

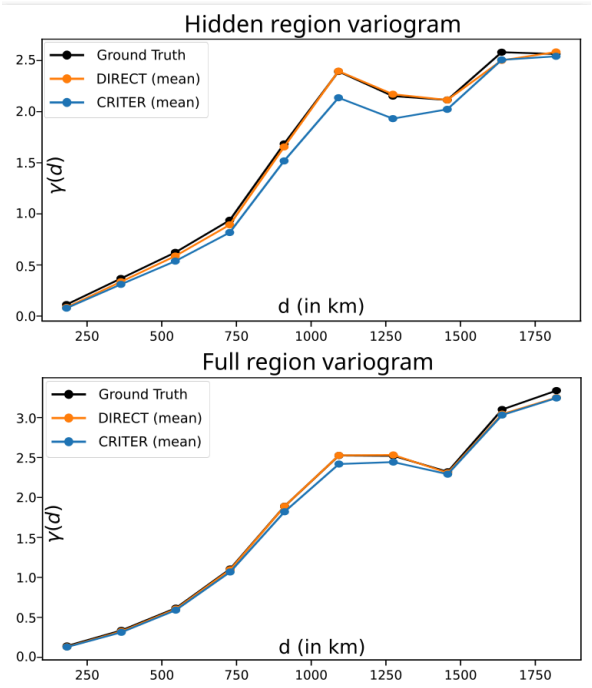
"[...] For each reconstruction, we first remove the large-scale spatial background by subtracting a least-squares linear plane fitted over the ROI, which suppresses the anisotropic low-wavenumber trend. We then compute the 2D gradient magnitude [...]"

"[...] Figures 10d and 11d show that both models reproduce the large-scale spectral structure of the SST field and follow the ground-truth spectrum closely from the largest scales down to intermediate wavenumbers. At higher wavenumbers both models depart from the ground truth and settle at similar energy levels. [...]"



Comment 5: Figs 12 & 13: the pixel scales are hard to interpret physically. Either replace with km or add an additional approximate km reference scale to help orient the reader.

Response: *As suggested, the spatial-lag axis in Figures 12 and 13 is now given in kilometers rather than in pixels.*



Comment 6: The CRPS values in Table 3 are not particularly informative without any form of reference to compare against. I'm not familiar with CRITER, but you present estimated uncertainty from that method so why can't we also get a CRPS for it? I'd honestly suggest just removing CRPS from the paper if there is nothing to benchmark against, another option would be to include MAE of the deterministic methods as reference values in the table since CRPS reduces to MAE for single member predictions?

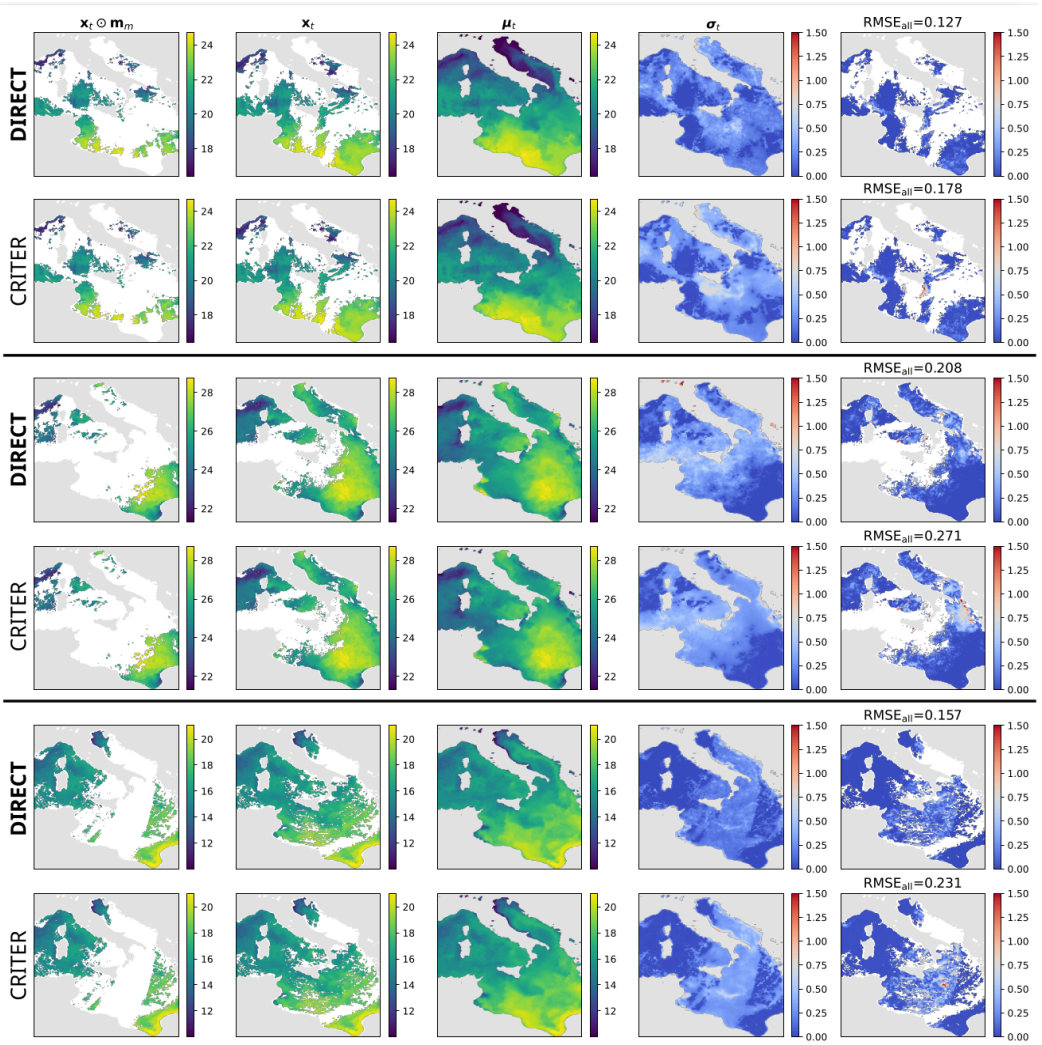
Response: *As suggested, we added CRITER as a baseline. Since CRITER predicts a per-pixel mean and variance rather than an ensemble, we compute its CRPS by sampling from the corresponding Gaussian. While adding this comparison, we identified and corrected an error in our original CRPS computation (an erroneous square-root). Section 4.2.6 and Table 5 are updated; DIRECT achieves lower CRPS than CRITER across all regions. The revised passage reads:*

> *Section 4.2.6 (CRPS):*
>
> *"[...] Unlike traditional deterministic error metrics, which assess only the mean reconstruction, the CRPS quantifies accuracy, statistical reliability, and sharpness of the predictive distribution. Rather than simply assessing whether DIRECT obtains a plausible distribution, the CRPS allows us to qualify if the produced ensemble actually explores the inherent uncertainty in the reconstruction and if probabilities derived from the ensemble are reliable. For DIRECT, the CRPS is computed empirically from the reconstruction ensemble. As CRITER does not produce an ensemble but instead predicts a per-pixel mean and variance, we compute its CRPS by drawing samples from the corresponding per-pixel Gaussian. Across all three regions, DIRECT achieves lower CRPS than CRITER, indicating that its ensemble provides a sharper and better-calibrated predictive distribution. As with the pointwise reconstruction errors, CRPS over the missing pixels is highest in the Atlantic region, reflecting the more extensive cloud cover and more energetic ocean variability."*

Dataset	Model	CRPS all [˚C]	CRPS mis [˚C]	CRPS obs [˚C]
Mediterranean	CRITER	0,052	0,172	0,014
	DIRECT	0,034	0,116	0,000
Adriatic	CRITER	0,049	0,132	0,012
	DIRECT	0,033	0,097	0,000
Atlantic	CRITER	0,199	0,316	0,037
	DIRECT	0,145	0,255	0,001

Comment 7: Multi-panel figures like Fig 6 throughout the paper would benefit from clearer sub-panel labels (e.g. panel a, b, etc.) and panel 6 is particularly hard to follow at first glance.

Response: *We revised the multi-panel comparison figures (e.g. Figure 6) by separating the per-day example blocks with a bold divider and emphasizing our model's name, so the two models can be compared directly on the same day. We considered explicit a/b/c sub-panel labels but found that, given the number of sub-panels, they made the figures more cluttered; the divider thus improves navigability while keeping the figures clean.*



Comment 8: Line ~109: typo "networ".

Response: *Thank you. We corrected the typo and have also proofread the manuscript and corrected the remaining typographical errors we identified.*

Comment 9: The fact the method can be trained exclusively on sparse real observations is to my mind a major strength compared to existing literature and could be better emphasized with a prominent mention somewhere in the abstract.

Response: We thank the reviewer for highlighting this point. We agree that training directly on sparse real observations, without requiring cloud-free or simulated ground truth, is a key strength of DIRECT relative to existing approaches. We have added a sentence to the abstract, making this explicit.

> *Abstract*
>
> "*[…] DIRECT is based on a rectified flow-matching formulation, conditioned on temporal context and day-of-year seasonality, and trained directly from sparse Level-3 observations using a self-supervised objective.*"

Reviewer #2:

This manuscript addresses an important problem and proposes a reliable method for reconstructing SST fields with uncertainty estimates. The experiments are extensive and generally convincing. I recommend a minor revision, with the following comments.

Response:
We thank the reviewer for the positive assessment and the helpful suggestions. We address each point below, indicating where the manuscript was changed and quoting the revised text.

Comment 1: The Introduction should include recent work on spatiotemporal diffusion models for climate and SST reconstruction, for example, Qian et al. (2026, arXiv:2602.16515) and Li et al. (2026, arXiv:2603.16272). A brief comparison would help clarify DIRECT's contribution relative to these studies.

Response: As suggested, *we extended the related-work paragraph in the Introduction to include both studies and compare them to DIRECT. For convenience, we reproduce the revised passage below:*

> *Introduction:*
>
> *"[...] Two concurrent studies extend this direction to global, climate-scale reconstruction: Qian et al., (2026) pretrain a spatiotemporal diffusion prior on climate-model simulations and reanalysis to reconstruct monthly historical temperature and precipitation fields from sparse instrumental records, while Li et al. (2026) perform diffusion posterior sampling to fuse coarse-resolution microwave retrievals with sparse in situ measurements into global ensemble SST estimates.*
>
> *While these works establish promise of generative modeling for geosciences, a crucial difference to the work presented in this paper is that they operate on a single shot or coarse temporal context (Choo et al., 2025; Qian et al., 2026), rely on guidance schemes without explicit temporal encoding (Song et al., 2025; Li et al., 2026), require gap-free training data (Qian et al., 2026; Choo et al., 2025) and do not target calibrated per-pixel uncertainty for dense spatiotemporal reconstruction (Wang et al. 2025)."*

Comment 2: Figure 2 and the corresponding description in Section 2 are somewhat confusing and difficult to follow. In particular, (x{t-1}) and (x{t+1}) are pixelwise composites from several dates, while (M_{t-1}), (M_t), and (M_{t+1}) have different meanings. A clearer schematic, example, or brief pseudocode would be helpful. The fixed three-day search range and nearest-valid-observation rule are also rather empirical, so their motivation and possible limitations should be discussed more explicitly.

Response: *We thank the reviewer. For clarity, we added Algorithm 1, which specifies the construction of the composite context frames and their temporal-origin masks step by step, and revised the surrounding text in Section 2 to separate the roles of the different masks. On*

the empirical design, we added a motivating sentence in Section 2 and discussed the limitation of the fixed window in the new Discussion section. The revised passage reads:

Section 2 (DIRECT architecture and implementation):

"[...] Analogously, missing pixels at $t+1$ are filled from $t+2$ or $t+3$. The resulting filled context frames are denoted as x'_{t-1} and x'_{t+1}. Note that this search depth $\Delta_{filled} = 3$ is distinct from the three-day context window, and extends the total temporal footprint of the input to days $t-3$ through $t+3$. A value drawn from an earlier day substitutes for the state at $t-1$ and stays informative only over the timescale on which SST remains temporally autocorrelated, so the same short temporal decorrelation that motivates a narrow context window also bounds the depth over which substitution is meaningful. To retain the temporal origin of each filled value, we associate every context frame with a one-hot encoded mask. For the past context, $M_{t-1} \in \{0,1\}^{3\times W\times H}$ encodes whether a pixel originates from day $t-1$, $t-2$, or $t-3$. For the future context, $M_{t+1} \in \{0,1\}^{3\times W\times H}$ similarly encodes offsets $t+1$, $t+2$, or $t+3$. Since each pixel is filled from the temporally nearest valid frame, these channels are mutually exclusive, while a null vector $[0,0,0]$ denotes a pixel that remains invalid (land or persistent cloud cover) after the search. The central frame is represented by a two-channel mask $M_t \in \{0,1\}^{2\times W\times H}$, which distinguishes observed from missing (or land) pixels at time t.

We define the temporal context fields as $C_{1t} = [x'_{t-1}, x_t, x'_{t+1}]$ and the corresponding masks as $C_{2t} = [M_{t-1}, M_t, M_{t+1}]$, whose construction for a general search depth is summarized in Algorithm 1. [...]"

Algorithm 1 Temporal context construction for target day t. The accumulator $\mathbf{f}$ marks pixels that have already been assigned a temporal origin, and $\mathbf{s}_j$ marks pixels filled from offset j. Because $\mathbf{s}_j$ is nonzero only where $\mathbf{f}=0$, the origin channels are mutually exclusive by construction.

Input: SST sequence $\mathbf{X}$, mask sequence $\mathbf{M}$, target day t, search depth Δ_{filled}
Output: context fields $\mathbf{C}_{1t}$, temporal-origin masks $\mathbf{C}_{2t}$

Past context:
$\mathbf{x}'_{t-1} \leftarrow \mathbf{0}, \quad \mathbf{f} \leftarrow \mathbf{0}$
for $j = 1$ **to** Δ_{filled} **do**
  $\mathbf{s}_j \leftarrow \mathbf{m}_{t-j} \odot (1-\mathbf{f})$ {pixels first filled from day $t-j$}
  $\mathbf{x}'_{t-1} \leftarrow \mathbf{x}'_{t-1} + \mathbf{x}_{t-j} \odot \mathbf{s}_j$
  $\mathbf{M}_{t-1}[j] \leftarrow \mathbf{s}_j$
  $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{s}_j$
end for
Future context: $\mathbf{x}'_{t+1}, \mathbf{M}_{t+1} \leftarrow$ as above, with $\mathbf{x}_{t+j}$ and $\mathbf{m}_{t+j}$
Central frame:
$\mathbf{M}_t[1] \leftarrow \mathbf{m}_t$ {observed}
$\mathbf{M}_t[2] \leftarrow 1 - \mathbf{m}_t$ {missing or land}
$\mathbf{C}_{1t} \leftarrow [\mathbf{x}'_{t-1}, \mathbf{x}_t, \mathbf{x}'_{t+1}]$
$\mathbf{C}_{2t} \leftarrow [\mathbf{M}_{t-1}, \mathbf{M}_t, \mathbf{M}_{t+1}]$

Discussion:

"The temporal context uses a fixed three-day window, following prior evidence that most relevant temporal information lies within a few surrounding days. This is limited when informative observations fall outside the window, most notably during cloud systems that persist over many days [...]"

Comment 3: Cloud masks from other dates are used to construct training samples. This is practical, but cloud occurrence may be coupled with atmospheric conditions and SST variability. Therefore, the missingness mechanism is not necessarily independent of the target SST field. The authors should discuss this assumption and clarify that the evaluation is based on synthetically withheld observed pixels rather than direct validation beneath real clouds.

Response: *We agree and discuss this point in the new Discussion section. The added text reads:*

Discussion:

"Both training and evaluation use cloud masks from other days, which enables learning without cloud-free ground truth. Two consequences follow. Cloud occurrence is coupled to the atmospheric state and hence to SST variability, thus the missingness is not strictly independent of the field being reconstructed. And because we synthetically withhold observed pixels, the evaluation measures reconstruction skill against a known reference rather than validating under genuine cloud conditions."

Comment 4: The variance correction provides a lightweight global calibration, but it does not necessarily demonstrate good pixelwise or conditional calibration. I suggest reporting the empirical coverage of nominal, for example, 50%, 80%, 90%, and 95% prediction intervals on the test set. Coverage before and after the correction would be particularly informative.

Response: *We thank the reviewer for the suggestion, which produced a useful insight. We now report empirical coverage of the nominal 50%, 80%, 90%, and 95% central prediction intervals over the withheld regions, before and after the correction, in a new table (Table 4) and expanded text (Section 4.2.3). As the reviewer anticipated, the global correction brings coverage close to nominal but does not achieve full conditional calibration, which we now state explicitly with its cause. For convenience, we reproduce the expanded passage below:*

Section 4.2.3 (Uncertainty estimation and bias analysis):

" We evaluate the reliability of ensemble-based uncertainty estimates over the withheld regions using two complementary measures: the scaled error ϵ of Eq. (5) and the empirical coverage of the nominal 50%, 80%, 90%, and 95% central

prediction intervals, where a withheld pixel is covered at level p when $|\epsilon| \leq z_p$ for the standard-normal quantile z_p. Results for the raw ensemble and after correction are given in Tables 3 and 4. Without the correction ($DIRECT_{w.o.\sigma_0}$), the ensemble is under-dispersed, with $\sigma_\epsilon > 1.7$ and coverage below nominal at every level and in all three regions. The two measures together indicate two distinct effects. Coverage is depressed across the full range of levels, including the central 50% interval, which reflects a broad under-dispersion of the ensemble. At the same time, the large σ_ϵ points to pronounced ensemble collapse at a sparse set of pixels, where the samples become nearly identical and the predicted variance is far too small. The learned regularization constant σ_0 (0.135 for Mediterranean, 0.101 for Adriatic, 0.298 for Atlantic), added in Eq. (3), addresses both effects. It raises the overall spread, correcting the global under-dispersion, and places a floor on the variance at collapsed pixels. With the correction, σ_ϵ falls close to one and coverage approaches nominal at the 80%, 90%, and 95% levels across all datasets. As a single global offset applied uniformly across pixels and probability levels, σ_0 does not achieve exact calibration everywhere: the corrected 50% intervals slightly over-cover (about 59-61 %) and the 95% intervals slightly under-cover (about 93-94 %), since one constant cannot adapt to spatially varying error regimes or to the tails of the distribution [...]"

Dataset	Nominal	Without sigma_0	With sigma_0
Mediterranean	50%	36.8%	58.7%
	80%	62.1%	82.0%
	90%	72.6%	89.1%
	95%	79.4%	93.0%
Adriatic	50%	36.5%	61.2%
	80%	61.8%	83.3%
	90%	72.5%	90.1%
	95%	79.5%	93.7%
Atlantic	50%	39.4%	58.6%
	80%	65.2%	83.2%
	90%	75.2%	89.8%
	95%	81.3%	93.1%

Comment 5: The satellite SST product is treated as ground truth, although it has its own retrieval uncertainty. I understand that explicitly incorporating observation errors would require substantial methodological changes, but this issue should at least be discussed because the reported residuals may include both reconstruction and observation errors.

Response: *We agree. Explicitly modeling observation error would require substantial methodological changes, we thus discuss the issue in the new Discussion section. The relevant text reads:*

> *Discussion:*
>
> *"DIRECT reconstructs an L3 satellite product, which is itself a retrieval carrying sensor noise and residual cloud contamination. Our metrics are computed against this product, so the reported residuals combine reconstruction error with the observation error already present in the reference [...]"*

Comment 6: I strongly recommend adding a short Discussion section covering the empirical temporal-context construction, the use of transferred cloud masks, the global nature of the variance correction, and the neglect of observation uncertainty.

Response: As the reviewer suggested, a Discussion section was added.