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.